

Multimodal Machine Learning for Precision Cancer Detection: Integrating Liquid Biopsy Biomarkers, Clinical Data, and Multi-Omics Profiles

 Shakhnoza Khamidova
Independent Researcher, USA

Received: 03 Dec 2025 | Received Revised Version: 14 Dec 2025 | Accepted: 23 Dec 2025 | Published: 31 Dec 2025

Volume 07 Issue 12 2026 |

Abstract

Accurate and early diagnosis of cancer is a significant challenge, as traditional diagnostic methods could be invasive, applicable only to specific cancer types, or inadequate to detect molecular changes in early disease. Liquid biopsy offers a minimally invasive, complementary source of circulating tumour DNA, cfDNA, RNA, and proteins, and multi-omics and clinical data offer complementary information about tumour biology and patient phenotypes. However, because of the heterogeneity and high dimensionality of these data, single-biomarker or unimodal analytical approaches are not very effective. Multimodal machine learning (ML) provides a framework to combine heterogeneous molecular, multi-omics, and clinical data to advance cancer detection and classification. In this review, we explore the conceptual framework and applications of multimodal ML to the field of precision cancer detection, focusing on the combination of liquid-biopsy-derived biomarkers, clinical features, and multi-omics data. Genomic, epigenomic, transcriptomic, proteomic, metabolomic, and fragmentomic features are discussed, and conventional ML, deep learning, and multimodal fusion strategies, such as early, intermediate, late, and hybrid integration, are evaluated. Applications involving early-stage cancer detection, cancer versus non-cancer classification, tissue of origin prediction, multi-cancer detection, and risk stratification are considered. The review further covers topics such as preprocessing, feature selection, model interpretability, model validation, integration of the model into clinical workflows, scalability, and the significant sources of biological, technical, computational, and dataset-related bias. While there is significant potential, there are still issues with limited cohorts, high dimensionality, missing modalities, generalizability, interpretability, privacy, and prospective clinical validation. The development of future progress will rely on the diverse multicentre cohorts, longitudinal monitoring, privacy-preserving learning, biologically informed and explainable learning, and prospective evaluation of clinical utility. Multimodal ML thus holds great potential for combining complementary information relevant to cancer and further developing more accurate, non-invasive, and clinically relevant cancer detection strategies

Keywords: Multimodal machine learning; precision cancer detection; liquid biopsy; circulating tumour DNA; artificial intelligence; early cancer detection; explainable AI.

© 2025 Khamidova, S. This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). The authors retain copyright and allow others to share, adapt, or redistribute the work with proper attribution.

Cite This Article: Khamidova, S. (2025). Multimodal Machine Learning for Precision Cancer Detection: Integrating Liquid Biopsy Biomarkers, Clinical Data, and Multi-Omics Profiles. The American Journal of Interdisciplinary Innovations and Research, 7(12), 150–167. Retrieved from <https://theamericanjournals.com/index.php/tajiir/article/view/8525>

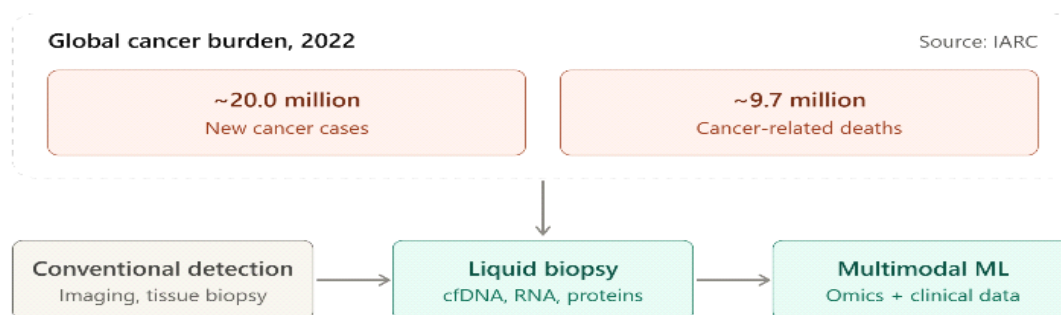
Introduction

Cancer is a significant health problem and has serious impacts on mortality, health care systems, and social and economic status around the world. In 2022, there were an estimated 19.98 million new cancer cases and 9.74 million cancer-related deaths worldwide, and there were an estimated 53.5 million people living within 5 years of a cancer diagnosis [1]. The global burden is likely to be further rising due to population growth, ageing, and the

changing exposure to cancer resulting in risk factors. Most importantly, the response to treatment and survival are strongly dependent on the stage of the cancer that is detected. Early diagnosis improves the likelihood of successful treatment and minimizes treatment-related morbidity and healthcare expenses [2]. Nonetheless, traditional screening and diagnosis are only available for specific types of cancer and can include imaging,

endoscopy, pathology, or tissue biopsy. However, screening can also yield false positive or false negative results, and

access to diagnostic infrastructure is not equal among populations [3].



• Global cancer burden and detection transition

Liquid biopsy has become a minimally invasive method of identification of molecular information originating from tumours in the blood and other body fluids. Complementary information on tumour biology can be provided by circulating cell-free DNA (cfDNA), circulating tumour DNA (ctDNA), circulating RNA, extracellular vesicles, circulating tumour cells, and proteins [4, 5]. Of these analytes, ctDNA can capture genetic and epigenetic changes whereas circulating RNA can reflect transcriptional activity and regulation. Proteins found in the circulation also carry extra information about pathways that could be involved in tumour development and the body's response to them. However, some of these biomarkers might have limited sensitivity (low signal in tumours), biological heterogeneity, variability in shedding of the biomarker, and limited specificity. Hence, a combination of several molecular signals may give a more detailed representation of the cancer-associated changes.

As molecules are becoming increasingly complex and large, the need for computers to recognise informative patterns in high-dimensional data have grown. Machine learning (ML) techniques can be used for feature selection, classification, dimensionality reduction and prediction, allowing complex molecular signatures to be linked to disease status. While challenges of data heterogeneity, limited cohort sizes, model interpretation and validation were still pertinent, the application of ML to multi-omics integration and clinical biomedical data had been growing [6].

Cancer is a complex and biologically variable disease and is not always simply characterised by a single molecular measurement. These different but complementary aspects of disease biology are captured by liquid-biopsy biomarkers, clinical characteristics, and multi-omics profiles. The multi-modal ML, therefore, can offer a paradigm to combine these disparate information sources into a single predictive model. In cancer detection, such integration could make it easier to detect cancer by combining molecular sensitivity with clinical context and multiple layers of biological information.

This review addresses the application of multimodal machine learning to address precision cancer detection, highlighting the fusion of the liquid-biopsy biomarkers, clinical data and multi-omics profiles. The review focuses on detection/diagnosis instead of treatment response/life course.

- **Conceptual Framework of Multimodal Machine Learning for Precision Cancer Detection**
- **Definition and Evolution of Multimodal Machine Learning**
- **Unimodal versus Multimodal Learning**

Traditional machine-learning methods for cancer detection typically process one type of data, like gene-expression patterns, circulating proteins, clinical features, or genomic changes [7]. While unimodal models can capture the pattern of disease in each dataset, they might not be a complete representation of the disease of cancer, which encompasses a variety of biological and clinical levels of complexity. Multimodal machine learning (MML) tackles this by combining data from multiple sources, but with

possibly complementary modalities, into a joint analytical model.

A major benefit of multimodal learning is that it allows for the acquisition of complementary, redundant and interacting information [8]. Circulating tumour DNA (ctDNA) can be used to obtain information regarding tumour-associated genetic and/or epigenetic alterations, while circulating proteins can also, to a certain degree, reflect functional changes and host responses. Additional phenotypic context is provided by clinical characteristics, which can impact the interpretation of molecular signals. By combining analysis of these information sources, machine-learning models have the potential to uncover relationships between them which cannot be identified when each information source is analysed individually [9].

There are also challenges to multimodal integration, though. Datasets can have significantly different dimensions, measurement scales, biological variation, missingness and technical quality. Hence, proper preprocessing, normalisation, feature selection and strategies for dealing with incomplete and heterogeneous data are important for effective multimodal learning.

- **Multimodal Learning Architectures**

There are basically two types of multimodal learning architectures, depending on the point in the learning process where the information from multiple modalities is fused together. Early fusion refers to the fusion of several modalities into one feature matrix prior to the training of a model. The strategy can be used to learn relations between variables between modalities but can be challenging to implement when the datasets differ significantly in dimensions or distribution.

The intermediate stages involve converting individual modalities into modality-specific representations and fusing modality-specific representations in a shared computational space. This method can maintain the properties of each feature class but also enable the model to learn relationships among classes. It is especially well suited to data with complex biological relationships in which relationships between molecular layers can be nonlinear [10].

Late fusion: Separate models are trained separately for each modality and then combined to obtain the prediction. This approach may be beneficial for patients for whom certain modalities are not available or for patients where the different types of data call for different analyses. Hybrid architectures integrate elements from early, intermediate,

and late fusion and may offer more flexibility for highly heterogeneous biomedical data sets.

- **Data Modalities Relevant to Cancer Detection**
- **Liquid Biopsy Biomarkers**

Liquid biopsy is a minimally invasive source of information on the molecular level and an important part of multimodal cancer diagnosis. Major analytes include cell-free DNA (cfDNA) and ctDNA, which can give information on genomic alterations and copy-number changes, DNA methylation and fragmentomic characteristics. Such characteristics have the potential to distinguish DNA from the tumour from the background cfDNA from normal tissues.

There is also molecular information in the form of circulating RNA. Alterations in the regulation of genes and cell activity related to malignancy can be reflected in microRNAs, messenger RNAs, long non-coding RNAs and other types of RNA. At the same time, circulating proteins can give information about signalling pathways that are associated with tumours, inflammation, tissue damage and host responses [11]. Extracellular vesicles can, in addition, carry bundles of nucleic acids, proteins and other molecules, which also convey molecular information.

Liquid biopsy would be an ideal approach for multimodal analysis given the complementarity of these analytes. Machine-learning models can combine multiple molecular signals into a more comprehensive cancer-like profile, rather than using a single biomarker.

- **Clinical and Demographic Data**

The clinical setting of a particular patient is a more informative context for molecular measurements. Biomarker distributions and cancer risk can be affected by demographic and clinical factors, including age, sex, smoking history, environmental exposures, symptomology, family history, laboratory parameters, and previous medical conditions. Phenotypic information can be added from imaging-derived characteristics [12].

Therefore, the combination of clinical variables and molecular biomarkers can help models differentiate the signals associated with cancer with the signals associated with age, lifestyle, inflammation, comorbidities, etc. Clinical information can also enhance individual risk estimation by adding contextual information not captured by molecular measurements.

- **Multi-Omics Profiles**

Multi-omics approaches offer complementary molecular information on multiple levels of biological organisation. Epigenomics is used to describe changes in the regulation of the genome, such as changes in methylation, and genomics is used to describe changes in DNA sequence. The transcriptomics analysis gives information about the activities of gene expression, while the proteomics analysis gives information regarding functional proteins and functional processes. Metabolomics reflects the changes in metabolites and biochemical pathways [13].

One layer of molecules per each molecular layer represents a different aspect of cancer biology. Genomic alterations can be the starting point or facilitator of malignant transformation, epigenetic modification change the regulation of genes, transcriptional changes are the result of altered cellular activity, proteins are functional consequences and metabolic changes are changes in cellular physiology. These layers can then offer a more complete representation of tumour biology than can analysis of a single omics dataset.

• Complementarity of Biological and Clinical Information

The overarching principles of multimodal cancer detection are linked to the fact that cancer is a complex and variable

biological process which may not be fully characterised by one biomarker or one type of data. Different aspects of the same disease process are described by the liquid-biopsy analytes, multi-omics profiles and clinical characteristics. The integration of these, therefore, provides a window of opportunity to correlate molecular changes with biological function and patient phenotype.

For example, a methylation or fragmentomic pattern observed in the plasma could be more diagnostic when combined with the clinical characteristics, protein profile, and smoking history of a single individual and correlated with the patient's age. Likewise, if changes are seen in both the DNA and RNA and protein layers, it is likely to be a stronger indicator of a disease-related process than a single molecular signal [14].

These relationships can be identified using "multimodal" machine learning, which offers a computational structure. A robust model should not only be able to learn the relationships among all these variables but also understand what information is complementary to each other, what signals are redundant or what features are most informative about cancer classification. This would help in making the detection models more robust, using many different biological and clinical aspects at the same time.



• Conceptual architecture of multimodal machine learning for precision cancer detection.

Multimodal cancer detection can be seen as a pipeline consisting of a few sequential steps: transforming heterogeneous patient information into harmonised

computational representation and then integrating patient information by an ML model, as shown in Figure 2. The goal is not just to expand the number of input variables but also to take advantage of complementary information from the molecular and clinical space to create a more

comprehensive and maybe more accurate cancer state representation.

• Liquid Biopsy Biomarkers for Machine Learning-Based Cancer Detection

Blood is carrying molecular information that is released by the tumour and the surrounding tissues, and liquid biopsy has become an important part of minimally invasive cancer detection. In contrast to traditional tissue biopsy, which only targets a small portion of a tumour's anatomy, liquid biopsy can offer a more representative picture of the

molecular changes in the tumour that are present in the bloodstream throughout the body. Rarely, however, an individual liquid-biopsy analyte is found at low levels and might be affected by biological or technical variation. A machine learning (ML) approach is available to extract complex patterns from such high-dimensional measurements and show that using multiple features of the biomarkers for cancer classification can improve the results [15].

- **Circulating Cell-Free and Tumour DNA**

Extracellular DNA fragments are called cell-free DNA (cfDNA), which are mainly released by apoptosis and necrosis of cells. Cancer patients have a certain percentage of cfDNA from tumour cells, known as circulating tumour DNA (ctDNA). The percentage of circulating tumour DNA (ctDNA) may greatly differ based on tumour type, disease stage, tumour burden, vascularisation and biological characteristics. This variability both poses a challenge and an opportunity for ML-based detection.

- **Genomic Alteration of Tumour**

Somatic mutations, copy-number alterations and structural variants are all genomic features of ctDNA. Mutations found in genes that are commonly found in cancer cells can be highly specific to the presence of cancer, especially if the gene is involved in the development of many kinds of cancer. Copy-number changes can offer data on genomic instability, and structural changes can detect rearrangements linked to specific malignancies [16].

Instead of relying on a single mutation, ML algorithms can be used to assess several genomic mutations at once. Feature-selection methods can select combinations of mutations or genomic patterns that are linked to cancer, while classification algorithms can leverage these patterns to classify cancer samples as cancer or non-cancer or to classify different cancer types. However, the sensitivity of mutation detection has been hindered by the low level of ctDNA present in the blood, especially in early-stage disease, where the proportion of tumour DNA in blood may be very low.

- **Epigenetic Features**

An additional source of information from cfDNA is the epigenetic alterations. One reason for the interest in DNA methylation patterns is the observation that the methylation abnormalities are very early events in the development of cancer and are linked to tissue-specific regulatory changes. Methylation profiling may assess more extensive regions of

the genome and may give information on tissue of origin and malignant transformation, unlike mutation-based approaches [17].

Other epigenetic characteristics, such as hydroxy methylation and chromatin-associated patterns, can also be involved in the detection of cancer. These datasets are ideal for ML methods, as the number of possible features from methylation profiles can be thousands. Algorithms can find combinations of differentially methylated regions or build predictive signatures of methylation to reduce the effect of irrelevant variables.

- **Fragmentomic Features**

In addition to sequence and methylation data, cfDNA also has information on DNA fragmentation. Fragment size distributions, fragment-end motifs, genomic coverage patterns and nucleosome-associated fragmentation can be indicators of the cellular process driving DNA release. Patterns of fragments across the whole genome could thus serve as indicators of cancer if certain mutations in the tumour are not easily identifiable.

Fragmentomic analysis is especially relevant for early detection since it can take advantage of high quality of many cfDNA fragments, not only rare sequence alterations specific to the tumour. The machine learning models can identify the fragmentation patterns that are associated with cancer from those that are not, using combinations of fragment size, genomic location and fragment-end characteristics.

- **Circulating RNA**

A complementary molecular layer is circulating RNA, which can be used to indicate changes in transcription and regulation that are linked to cancer. Multiple classes of RNAs have been studied as possible liquid biopsy markers.

- **miRNA**

The microRNAs (miRNAs) are short non-coding RNAs which regulate gene expression at the post-transcriptional level. miRNAs have been found to be dysregulated in various types of cancer and can be identifiable in the bloodstream because of binding to proteins, lipoproteins or extracellular vesicles. Thus, miRNA signatures can be created that are associated with cancers.

Multiple miRNAs can be analysed concurrently using ML methods, and combinations of miRNAs can be found that offer a higher discriminatory capacity than the individual miRNAs. Differences in miRNA expression could also be

used to differentiate cancer from benign conditions or to classify different types of cancer in such models. Circulating miRNA measurements may, however, vary depending on the tissue, physiological status, and pre-analysis variables [18].

- **mRNA and Long Non-Coding RNA**

Circulating messenger RNA (mRNA) may be used for information on transcriptional activity, and long non-coding RNA (lncRNA) is involved in a variety of regulatory mechanisms related to the development of tumours. Their expression patterns may then offer other clues to help classify cancer. The expression signatures of a multigene set can be retrieved from expression measurement of circulating RNA molecules by ML, which may not be discernible from the expression of individual RNA molecules.

- **Other Circulating RNA Species**

Additionally, other types of non-coding RNA that have emerged for interest due to their stability and regulatory role are circular RNAs (circRNAs). Integration with existing RNA and DNA biomarkers can improve the molecular landscape of liquid biopsies. Before these new classes of RNA can be routinely used in clinical workflows for detection; however, standardised detection approaches and larger validation cohorts are needed.

- **Circulating Proteins**

- **Protein Biomarkers**

Proteins in the circulation deliver information which is different from the information in the nucleic-acid categories of biomarkers. Proteins that are secreted by tumours can be indicative of changes in signalling, damage,

and interactions with the host. Furthermore, inflammatory and immune proteins can also trap systemic inflammatory and immune responses related to malignancy.

Due to the limited specificity of individual proteins, ML can measure more than one protein concentration at a time, looking for combinations that are linked to cancer. This approach may allow the identification of signals derived from the tumour and from the host to distinguish between malignant and non-malignant disease.

- **Protein Panels and Multiplex Measurements**

Numerous circulating proteins can be measured simultaneously on multiplex protein platforms. Then, using ML algorithms, informative subsets can be identified, and multi-protein diagnostic signatures built. These signatures can offer enhanced discrimination relative to single-marker signatures due to the ability to capture multiple biological processes at the same time. Protein levels, however, can be affected by inflammation and other physiological factors, such as age and comorbidities, so careful cohort selection and model validation are crucial.

- **Strengths and Limitations of Individual Liquid-Biopsy Modalities**

All the liquid biopsy techniques offer a different aspect of cancer biology. Circulating DNA (ctDNA, cfDNA) gives information on genomic, epigenetic and fragmentomic information; circulating RNA gives information on transcriptional activity and regulation; circulating proteins give information on functional and host response. They have complementary properties, which are well suited for multimodal ML that allows multiple molecular layers to be combined and not analysed separately [19].

1. *Liquid-biopsy biomarker classes and ML-relevant features.*

Biomarker	Molecular information	Typical features	Advantages	Limitations
ctDNA	Genomic/epigenetic	Mutations, methylation, copy-number alterations	Tumour-associated molecular signal	Low abundance, especially in early disease
cfDNA	Genomic/fragmentomic	Fragment size, end motifs, genome-wide fragmentation	Broad molecular coverage	Substantial non-tumour background

miRNA	Regulatory	Expression profiles and signatures	Relatively stable in circulation	Tissue specificity and biological variability
mRNA/lncRNA	Transcriptional/regulatory	Expression signatures	Functional and regulatory information	Stability and pre-analytical variation
Proteins	Functional/host response	Concentrations and multiplex panels	Clinically accessible and measurable	Limited specificity of individual markers

The potential of liquid biopsy is then not only to detect a single molecular signal but rather to collect complementary signals. The computational structure needed to combine all these heterogeneous features and identify complex biomarker patterns related to early cancer diagnosis can be provided by multimodal ML.

• Multi-Omics and Clinical Data Integration

Cancer is a complex process involving interwoven changes at several levels of biology that are hard to represent with any single molecular dataset. Each of the four domains (genomics, epigenomics, transcriptomics, proteomics, and metabolomics) of the 'omics' has unique information regarding disease pathogenesis and patient phenotype. Combining these complementary data types with multimodal machine learning (ML) can provide a more complete picture of cancer and enhance the accuracy of detection. But it is important to account for variations in data structure, scale, quality, and biological significance when integrating [20].

- Genomics and Epigenomics
- Mutation and Copy-Number Profiles

Genomic changes are the basic tools in understanding the molecular basis of cancer. Abnormalities that can be detected by somatic mutations, copy-number alterations and structural changes are associated with malignant transformation. If, on the other hand, the genomic information comes from liquid biopsy, the DNA of the tumour may be at low levels in a large background of DNA from non-cancerous cells. This can decrease the sensitivity of the individual genomic markers, especially in early disease.

These can be overcome by machine-learning-based methods, which can consider several genomic attributes in a single analysis. Models can detect combinations of mutations or copy-number patterns or broader genomic signatures linked with cancer, rather than relying on a single mutation. Yet, genomic profiles can be combined with other molecular and clinical data to enhance the information gained.

• DNA Methylation Signatures

Epigenomic changes are another important source of cancer information. Abnormal methylation in tumour development can affect the regulation of genes which are linked to proliferation, differentiation and tumour suppression. Thus, methylation patterns may be diagnostically useful even in the presence of a poor genetic profile of the tumour.

A genome-wide methylation profile generates a tremendous number of potential features and hence is ideal for computational investigation. ML algorithms can detect informative methylated regions, reduce the dimensionality of features and create prediction methylation signatures. Further, tissue-specific methylation signatures may be used to help determine the source of the cancer signal, adding another layer to the multimodal cancer detection.

- Transcriptomics
- Gene-Expression Profiles

Transcriptomic data reflect the activity of the genes and are a functional level between genetic changes and downstream cellular function. These changes in the expression of genes may be due to changes in signalling pathways, cell proliferation, immune functions, and

interactions between the tumour cells and the surrounding tumour microenvironment.

Transcriptomic data is very high dimensional and poses a challenge and opportunity for ML. While thousands of genes can be measured at once, just a few may have useful diagnostic information. Feature-selection and dimensionality reduction methods can then be used to reduce the risk of overfitting while identifying informative expression patterns. To link changes in structure with their consequences, a combination of gene-expression profiles, genomic and epigenomic data could be used as well [21].

- **Non-Coding RNA Signatures**

Non-coding RNAs (ncRNAs), such as microRNAs, long non-coding RNAs, and circular sRNAs, play roles in regulating gene expression and cell behaviour. They have been linked to various cancer-related events and could be considered as part of cancer molecular diagnostic signatures.

Unlike evaluating each individual RNA molecule, ML can look for coordinated expression patterns of multiple non-coding RNAs. Combined with the integration of coding-gene expression and other molecular datasets, these signatures could yield further information on regulatory alteration in malignancy.

- **Proteomics and Metabolomics**
- **Proteomic Signatures**

Proteomic profiles offer data on proteins that participate in the cells' signalling, metabolic, immune, and other functional processes. Proteomes contain downstream products of genomic and transcript activity, which means that proteomic information can complement DNA- and RNA-based measurements.

Multiplex protein profiling provides the ability to assess many blood or tissue proteins at a time. In addition to identifying combinations of proteins linked to cancer, ML can also decrease the need for individual proteins that could be specific for only a subset of cancers. Combination of proteomic data and genomic and transcriptomic data can help to bridge the gap between changes at the molecular level and their functional outcomes [22].

- **Metabolic Alterations**

Metabolomics yields information regarding small molecules that play a role in cellular metabolism and

regulation of the physiology. Increased proliferation, metabolic changes, changes in energy utilisation, biosynthesis, and interaction with the tumour microenvironment are hallmarks that are often observed in malignant cells. These changes may produce a metabolic signature that complements other molecular layers.

One of the most useful applications of metabolomics is its integration with transcriptomics and proteomics data, which can be useful for understanding how molecular changes at the genome and transcript level can have downstream effects at the metabolic level. There are, however, other factors that can be affected by metabolomic measurements, such as diet, medication, lifestyle, physiological state, and sample-handling conditions, which require appropriate standardisation [22].

- **Clinical and Demographic Information**
- **Patient-Level Characteristics**

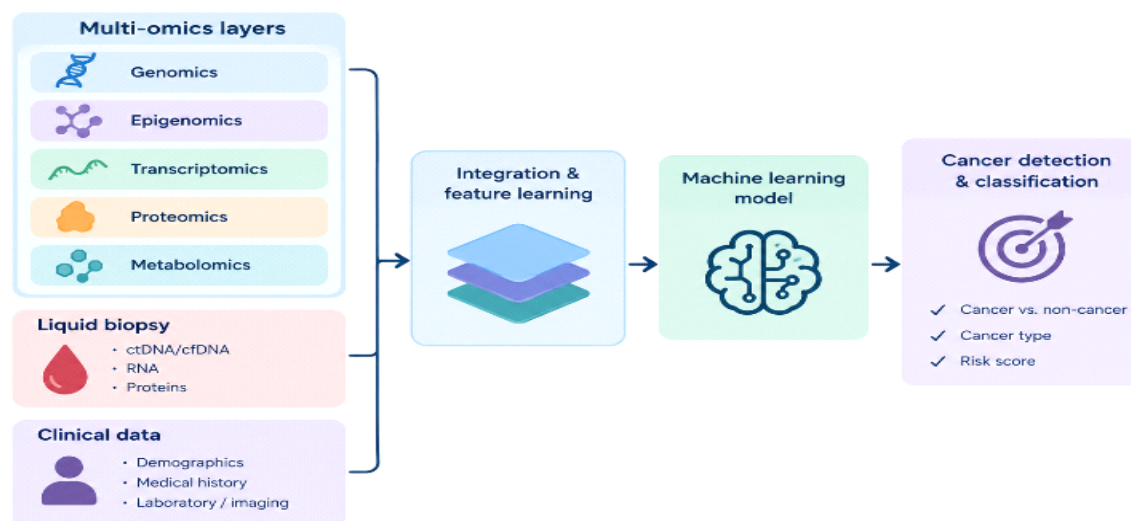
The clinical and demographic data offer a background to understanding the molecular measurements. Cancer risk and biomarker profiles can be affected by factors like age, sex, family history, symptoms, past health conditions, and past disease history. These properties can thus be incorporated into ML models and aid in patient-specific interpretation of molecular signals.

- **Laboratory and Physiological Measurements**

Other laboratory tests may be done routinely to give more information about systemic changes that may occur in cancer. Physiological variables such as haematological parameters, inflammatory markers, biochemical measurements, and other variables can be related to processes associated with the tumour or to mechanisms of the host response [23]. These features, together with molecular biomarkers, may have the potential to enhance malignant versus non-malignant discrimination.

- **Environmental and Lifestyle Factors**

Factors in the environment and lifestyle of the person, such as exposure to tobacco, alcohol, work exposures, diet, and other behavioural factors, can affect cancer development and molecular profiles. They can assist in determining the baseline risk in an individual and can be used to put molecular abnormalities in context. This is especially significant if the biomarkers are affected by something other than malignancy.



- **Incorporating biological information layers with multimodal machine learning.**

The various layers of data can be represented as a biological hierarchy, from the genomic and epigenomic data to transcriptional, protein, and metabolic changes to the clinical phenotype (see Figure 3). Liquid biopsy fragmentomic information can offer an added molecular layer between the circulating DNA and the downstream biological properties. These layers can be combined in a single representation by using multimodal ML so that the relationships between the molecular changes and the characteristics at the patient level can complement each other in the detection of cancer.

- **Challenges in Combining Heterogeneous Data**

Though there are potential advantages for the integration of multiple omics datasets, there are several challenges in terms of both technique and biology. Different modalities have different measurement technologies and scales. For instance, genomic data might have thousands of molecular variables, while clinical data might have only a small number of variables of a demographic or physiological nature. High-dimensional modalities can be more influential on model development without proper normalisation and preprocessing.

Another significant concern is missing data, as patients may not have measurements from all molecular modalities.

The reasons why data might be missing include inadequate sample size, assay failure, price, or study design. Such models may, therefore, be less applicable to real-world clinical settings if they do not allow for the inclusion of missing data.

The sources of batch effects are sample preparation, laboratory procedures, sequencing platforms, measurement technologies, or processing pipelines. These technical variations can be learnt rather than disease related signal, if not properly controlled in an ML model. Likewise, the feature selection may be unstable, and the risk of overfitting is high due to the large number of features compared to the number of samples for multi-omics data.

Temporal differences need to be considered as well. Molecular and clinical measurements taken at various times could be different states of the disease, especially if it changes. There are also inconsistencies between clinical definitions, annotations may not be complete and there may be variation in the characteristics of the cohorts.

Thus, for proper multimodal data integration, meticulous quality control, normalisation, managing batch effects, feature selection, missing data management and suitable validation are necessary. The goal is not just to include as many variables as possible but to retain useful information from each modality and enable the ML model to establish meaningful relationships between complementary variables that aid accurate cancer detection.

- **Machine Learning Strategies for Multimodal Cancer Detection**

Such multimodal cancer detection approaches are not only heavily reliant on the type of machine-learning algorithm used but also on the proper preprocessing, feature representation, integration strategy, and the interpretation of the model. The data from the liquid biopsy, clinical studies and multi-omics studies vary greatly in scale, dimensionality, distribution and biological meaning, and careful data preprocessing is needed before integration [24].

- **Data Preprocessing and Feature Engineering**
- **Quality Control and Normalisation**

Quality control gets unreliable molecular measurements, technical artefacts, outliers, and poor-quality samples that could lead to false signals. Afterward, measurements are normalised to compare them between samples and platforms. Because of variations in laboratory methods, sequencing technologies, reagent lots, or acquisition dates, multi-omics datasets can introduce systematic variation that is not cancer-related, especially when analysing these types of batch-corrected datasets.

- **Missing-Data Management**

Multimodal datasets often contain missing data, as it is not possible for all patients to receive all molecular tests. Simple imputation might be sufficient for small amounts of missingness, while model-based or modality-aware imputation can be employed for patterns of missingness. Importantly, the absence of a measurement should not be assumed to be random, as the missingness might be caused by clinical or experimental parameters.

- **Feature Selection and Dimensionality Reduction**

There can be thousands of molecular variables in liquid biopsy and multi-omics datasets for each patient available. Feature selection techniques help to find informative biomarkers, and dimensionality reduction techniques help to extract lower-dimensional representations of highly correlated variables. These steps are designed to minimise overfitting of the data, the computational load and noise, and to make the model more generalisable.

- **Conventional Machine Learning Approaches**

The traditional algorithms are still useful for the detection of cancers across multiple modalities, especially when dealing with a small data set. Logistic regression and

regularised models can efficiently perform classification and can discover combinations of predictive biomarkers. SVMs are good in high-dimensional feature spaces, while decision trees and random forests can be used to capture nonlinear relationships and interactions. The gradient boosting methods gradually grow accurate predictors by aggregating weak ones to build strong learners. These are summarised in Table 3, along with the main characteristics and possible applications of such approaches [25].

- **Deep Learning Approaches**

Complex nonlinear representations can be learned directly from high-dimensional molecular and clinical data using deep-learning methods. The ability to model interactions among heterogeneous features and the ability to obtain local patterns from structured molecular representations are modelled by artificial neural networks and convolutional neural networks, respectively. Autoencoders are valuable for dimensionality reduction and learning latent biological representations. Other representation-learning methods can also represent sequential or longitudinal patient data. But the disadvantage of deep models is that they need a large amount of data, and their results are hard to explain compared to standard models.

- **Multimodal Integration Strategies**

The multimodal integration can happen at various levels of the analytical process. Early fusion combines features from liquid biopsy, multi-omics and clinical data into one vector of features which is later used for training the model. This is a simple-to-implement method that may be sensitive to dimensionality and missing modalities.

For intermediate fusion, modality-specific encoders are used, and, subsequently, the modality-specific representations are mapped to a common latent space. The strategy can be used to preserve modality-specific information and learning relationships between different biological layers.

In the late fusion approach, different models come up with their own predictions, and the predictions are fused together with votes, weighted averages, stacking, or any other fusion methods. This is helpful when the structure of the modalities is different or when a patient does not have some of the measurements. The three major strategies are conceptually compared in Figure 4.

A. Early fusion**B. Intermediate fusion****C. Late fusion**

 Stage where modalities are combined

- Multimodal ML integration strategies.**

Hybrid integration can combine these strategies, for example, by learning modality-specific representations before jointly integrating selected features and final predictions.

- Model Interpretability and Explainable AI**

Predictive performance is not sufficient to define the bio-clinical usefulness, and interpretability is critical for clinical application. Feature-importance methods can indicate the biomarkers that are particularly influential, and

SHAP-type approaches can estimate the impact of features on individual predictions. Deep-learning models can have attention mechanisms that identify informative regions or variables in the molecule. Biological interpretability could be achieved by determining whether selected features refer to biologically relevant cancer pathways, while clinical interpretability could be achieved by making predictions related to patient characteristics and clinical decisions. Overall, explainable AI can assist in the discovery of biomarkers, the validation of models and the confidence of the clinicians in multimodal cancer detection systems [26].

2. Machine-learning algorithms applicable to multimodal cancer detection

Algorithm	Data type	Main strength	Major limitation	Interpretability	Potential application
Logistic/regularised regression	Molecular + clinical	Simple, efficient	Limited nonlinear modelling	High	Biomarker-based classification
SVM	High-dimensional molecular data	Effective with small datasets	Parameter-sensitive	Moderate	Cancer classification
Random forest	Mixed/tabular data	Handles nonlinear interactions	Less transparent with many trees	Moderate	Biomarker and risk prediction
Gradient boosting	Tabular multimodal data	Strong predictive performance	Can overfit without tuning	Moderate	Diagnostic classification

Neural networks	High-dimensional multimodal data	Learns complex representations	Data-hungry, less interpretable	Low-moderate	Multimodal cancer detection
Autoencoders	Multi-omics/molecular data	Latent feature learning	Latent features may be difficult to interpret.	Low	Dimensionality reduction

• Applications of Multimodal Machine Learning in Precision Cancer Detection

Recent years have seen the increased use of multimodal machine learning for cancer detection, leveraging molecular biomarkers with additional clinical and phenotypic data. These methods can use a combination of circulating DNA, RNA, proteins, multi-omics profiles and patient features to enhance the ability to discriminate malignant from non-malignant states, rather than being restricted to a single biomarker [27].

- **Early Cancer Detection**
- **Detection of Early-Stage Disease**

In early-stage tumours, levels of blood biomarkers are low and may not be detected. The combination of genomic, epigenetic, fragmentomic, transcriptomic, and protein features in multimodal models can overcome the weak signals in individual models. Integration can then help to identify subtle molecular changes in early tumour development and decrease reliance on a single highly sensitive biomarker.

- **Cancer Versus Non-Cancer Classification**

One of the major applications is to differentiate cancer patients from normal patients or patients with benign diseases. The amalgamation of liquid-biopsy measurements and clinical factors can be complementary to the ability of models to detect molecular patterns that are linked to malignancy while simultaneously capturing patient-level features. This is especially applicable to screening, where high specificity is important to reduce unnecessary further investigative work.

- **Detection Across Multiple Cancer Types**

Multimodal methods can be used to facilitate detection of multiple cancer types in a common analytic framework. The ability to use common molecular signals and cancer-specific characteristics could allow models to differentiate

malignant from non-malignant samples while also providing the likely cancer type. This provides a basis for multi-cancer early-detection strategies.

- **Cancer-Type Classification**
- **Tissue-of-Origin Prediction**

If a cancer signal is found, it is important to find out which tissue the cancer originated from. Circulating DNA, methylation, RNA and proteins have tissue-associated signatures at the molecular level. These features can be incorporated into machine-learning models for estimation of the most likely anatomical source, which could help inform diagnostic imaging and tissue-directed evaluation.

- **Multi-Cancer Detection Approaches**

The goals of multi-cancer detection models are to detect multiple cancer types with the use of a single blood-based assay. The accuracy of cancer detection and the correct classification of the cancer tissue of origin are the factors influencing their performance. Possibly by integrating multiple molecular signals, rather than a single class of biomarkers, multimodal integration has the potential to enhance this process.

- **Risk Stratification and High-Risk Populations**

Multimodal ML can also be applied to people who are at higher risk for cancer. The individual probability of cancer can be estimated by the combination of age, family history, previous disease, environmental exposure, lifestyle factors and molecular biomarkers. These models could potentially be used to prioritise people for more testing in high-risk groups, while exposure-associated features may increase the likelihood of risk assessment in groups exposed to tobacco, workplace risks, or other cancer-causing substances.

- **Integrating Multimodal Biomarkers with Clinical Phenotypes**

The clinical variables can be crucial in the context of the molecular signals. The use of a combination of biomarkers in addition to symptoms, demographics, laboratory, and clinical history may enhance sensitivity by capturing patients with low molecular signals and clinically suspicious histories. Concurrently, the inclusion of clinical context could help to increase the specificity because it can distinguish cancer-related molecular changes from those that occur during benign and inflammatory processes. This cohesive approach can minimise misclassification and provide a more individualised diagnostic probability, not just a binary prediction.

- **Comparative Performance of Unimodal Versus Multimodal Models**

Unimodal models tend to be simpler to construct, and they can work well when there is a strong disease signal in the biomarker. But they may be restricted due to biological

heterogeneity, low tumour size or non-specific molecular changes. The multimodal models can integrate complementary information from the molecular and clinical space and can potentially yield improved classification. These benefits, however, should be shown to be present in properly designed validation sets, as the introduction of a modality does not necessarily lead to an improvement in generalisation. Meanwhile, the observed performance of these cohorts may vary significantly depending on the underlying differences in the cohorts, as well as differences in preprocessing, feature selection, and evaluation metrics.

The applications of liquid biopsy and multimodal machine learning are summarised in the literature by representative applications as shown in Table 3. The table highlights the type of cancer, the characteristics of the cohort, the types of biomarkers evaluated, the analytical approach, and the main findings, with a particular focus on limitations that could influence clinical interpretation.

3. Representative applications of AI/ML in liquid biopsy and multimodal cancer detection

Study/Application	Cancer type	Data modality	ML approach	Main outcome	Key limitation
cfDNA/ctDNA analysis	Multiple	Mutations, CNAs, fragmentomics	ML classifiers	Cancer detection	Low tumor signal
DNA methylation	Multiple	cfDNA methylation	RF, SVM	Detection & tissue-of-origin	Biological variability
Circulating RNA	Multiple	miRNA, mRNA, lncRNA	SVM, RF, neural networks	Cancer classification	RNA instability
Protein panels	Multiple	Circulating proteins	Regression, ensemble ML	Improved diagnostic discrimination	Limited specificity
Multimodal liquid biopsy	Multiple	DNA RNA/protein	+Multimodal/ensemble models	Improved integrated prediction	Larger validation cohorts needed
Biomarker + clinical data	Multiple	Molecular + clinical variables	+ML/deep learning	Personalized risk prediction	Missing/heterogeneous data

- **Clinical Translation and Precision Oncology**

Multimodal machine-learning models must be accurate in predictions, but needing to move them from the lab to the clinic is an undertaking more than a simple matter of accuracy. The system should perform well in the clinical laboratory, show clinical validity, be clinically useful, reproducible and compatible with the current clinical procedure [28].

- **From Computational Model to Clinical Diagnostic Test**
- **Analytical Validity**

Analytical validity is the ability of the entire testing process to properly and reliably measure the desired biomarkers. This involves a standardised sample collection method, a standardised molecular extraction, standardised sequencing or measurement of the protein, quality control and computational processing for multimodal liquid-biopsy systems.

- **Clinical Validity**

Clinical validity establishes the reliability of the association between the molecular and clinical characteristics that are discovered by the model with the presence or absence of cancer. Appropriate measures of performance (sensitivity, specificity, predictive values, and discrimination) should be used in appropriate patient populations.

- **Clinical Utility**

The clinical utility should be assessed as to whether the outputs of the model enhance real-world clinical decision-making. Ideally, a multimodal test should act as a discriminator for patients who need to go for further testing, be a risk factor for over triage, and not generate a high number of false-positive results.

- **Model validation and generalisability**

Strong validation is critical prior to clinical use. External validation can be used to test model performance in a different sample of patients, whereas internal validation can be used to test the stability of the model within the development sample. Cohorts working in multiple centres and across a wide geographical area are of special importance in determining if models are generalisable beyond the population on which they were trained. Prospective validation offers greater rigour in the evidence of real-world use, as the samples and decisions are made in the real world and not selected in retrospect.

- **Clinical Workflow Integration**

The system should include the clinical deployment of the laboratory processing, molecular profiling, pre-processing, multimodal prediction and clinical interpretation steps in a reproducible workflow. The automated calculation pipelines can minimise the variability of processing, and the interfaces to make decisions can show model predictions together with relevant biomarkers and clinical information. Integration should be used in addition to clinical judgement, not in place of it.

- **Cost, Scalability and Accessibility**

Sequencing, molecular assays, laboratory infrastructure, computational resources, and personnel should be considered when implementing at a large scale. Although high throughput and automated analysis can increase the scale, complex multimodal testing can be expensive. Simplification of biomarker panels and the efficient use of computational architectures could make it easier to implement across a wider range of settings. This summary of the overall translation pathway from patient sampling to clinical decision-making is presented in Figure 5.

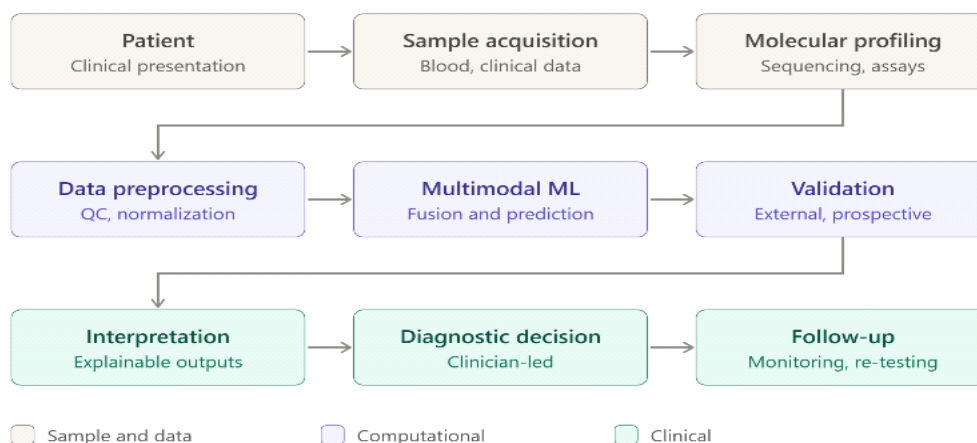


Figure 5. Multimodal cancer detection clinical pathway:

Scalability should be thought of alongside the diagnostic accuracy, affordability, accessibility and ability of healthcare systems to deal with people who do not have Covid, but who test positive or have an indeterminate result.

• Challenges, Limitations and Sources of Bias

Although multimodal machine learning (MML) has great promise for precision cancer detection, several biological, technical, computational, and clinical challenges can impact the reliability of models. These challenges need to be addressed to develop accurate, repeatable, clinically transferable systems [29].

• Biological and Technical Heterogeneity

There are significant differences in the genetic, epigenetic, transcriptional and metabolic properties of tumours. This variability can manifest itself in varying biomarker signatures even within a cancer type. The overall circulating tumour signal may also be contributed by another source of material, such as material from the host, in early disease. Additionally, temporal fluctuations in biomarker measurements may occur due to biochemical changes within the tumour or due to a change in the status of the therapy. Furthermore, there are pre-analytical variations due to differences in blood collection, processing, storage and laboratory procedure.

• Computational Challenges

Multimodal datasets typically have many thousands of features related to the molecules and a relatively small number of patient samples. This is a small sample problem with a lot of dimensions which will heighten the chances of overfitting. Another type of class imbalance may be when the ratio of cancer and control is skewed or when specific cancer types don't have enough cases. The lack of modalities causes further challenges, as not all patients might have genomic, transcriptomic, proteomic and clinical data. Model learning can be further skewed by batch effects and inter-platform variability. Analytical platform differences and batch effects can further skew model learning.

• Dataset and Cohort Limitations

Most studies are based on retrospective or single-centre data, which might not be sufficient to represent real-world populations. Generalisability may be limited due to lack of geographic, ethnic, demographic and clinical diversity. One source of selection bias is when patients in the analysis differ systematically from those in the population to be screened; one source of spectrum bias is when the assessment of the models is based on an analysis that included cancer and healthy control groups with very clear clinical distinction between the groups.

• Model-Related Bias and Reproducibility

Often, data leakage, inappropriate features and too small of a separation between the train and test sets can result in inflated performance estimates. Comparisons between studies are also hindered by differences in the

preprocessing procedures, feature definitions, evaluation measures, and validation approaches. Reproducible pipelines, clear reporting, independent test groups, and standardised performance metrics are thus imperative.

• Interpretability and Clinical Trust

Complex multimodal and deep learning models can also be “black boxes”, meaning it may be hard to understand why a specific patient is considered high or low risk. Clinicians are seeking predictions that are statistically reliable, biologically plausible and clinically meaningful. Explanatory AI can be used to uncover influential biomarkers and give interpretable evidence to back up the model predictions.

• Regulatory, Ethical and Privacy Considerations

Sensitive clinical and genomic information might be stored in multimodal systems, which may pose significant data protection and responsible data governance needs. One-way algorithmic bias can occur if the training data doesn't sufficiently include certain populations. Several regulatory, data management, reporting, and performance monitoring requirements are necessary to enable this clinical application, including proper regulatory controls, safe data handling, clear reporting, and monitoring model performance.

4, Major challenges and potential solutions for multimodal ML

Challenge	Consequence	Potential solution
Small cohorts	Overfitting	Larger multicentre cohorts
High dimensionality	Unstable models	Feature selection and regularisation
Missing modalities	Reduced applicability	Imputation and flexible models
Batch effects	Technical bias	Standardisation and batch correction
Dataset bias	Poor generalisability	Diverse external cohorts
Black-box predictions	Limited clinical trust	Explainable AI
Data leakage	Inflated performance	Strict independent validation

• Conclusion

Multimodal machine learning is shifting from monomarker to multimarket-based approaches to cancer detection, which integrate molecular, clinical and phenotypic data. There will be increasingly more dynamic, comprehensive, interpretable and clinically integrated systems in the future. One of the key directions is longitudinal multimodal monitoring, where periodic sampling of the liquid biopsy allows tracking of temporal dynamics of the level of circulating DNA, RNA, proteins, and other biomarkers. These might enable the acquisition of data not just for detection but also for disease evolution, response to treatment and disease recurrence.

Distributed clinical and genomic data resources are growing, providing opportunities for distributed learning

and privacy-preserving learning. Institutions would be able to create and test models together without sharing sensitive patient information. This can help to create bigger and more varied training groups and resolve key issues of genomic data privacy and governance.

Other bimodal systems that could be included in the future are liquid biopsy, multi-omics, medical imaging, electronic health records, and digital-health data. Hyperspectral imaging of this complementary information sources may be used to represent a more complete image of tumour biology and patient phenotype. But powerful approaches to integrating the differences in analytical platforms, the missing modalities, temporal information, and heterogeneous data structures will be essential.

There should also be increased focus on explainable AI and biologically inspired AI. The addition of molecular pathways, biological networks, and mechanistic relationships can enhance the interpretability of model predictions and aid in the discovery of biomarkers. Clinically interpretable outputs will be of special value in helping physicians understand the reasons for a model's classification of a patient as high risk or low risk and how that classification should affect the diagnostic decisions he or she makes next.

In the end, future clinical trials, multi-centre studies testing the work, and real-world testing will determine the success of the translation. Future investigation needs to evaluate clinical utility, cost effectiveness, workflow, availability, and patient outcomes, as well as sensitivity and specificity. To identify if the models created in the selected research populations are also valid in other clinical, demographic, geographic, and screening populations, diverse populations need to be screened.

Finally, multimodal machine learning systems hold great promise for the early detection of cancer with a high degree of precision, combining complementary information derived from liquid biopsy biomarkers, multi-omics profiles and clinical data. Together, liquid biopsy provides molecular information with minimal invasiveness, multi-omics broadens the scope of biological information, and clinical variables offer patient-level contextualisation. Integrated forms of their use have the potential to enhance early cancer detection, cancer-type classification, risk stratification, and individualised diagnostic assessment; when used alone, cancer diagnosis is limited to using a single biomarker.

Despite this, there are many challenges that remain, such as biological and technical diversity, high dimensionality, missing data, limited diversity within cohorts, overfitting, algorithmic bias, poor external validation, lack of interpretability, privacy concerns, and poor clinical standardisation. To overcome these shortcomings, strict pretreatment, an open training set, multicentre datasets, explainable methods, and prospective validation are needed. Thus, the true power of multimodal AI will require not just high predictive accuracy but also the ability to show clinical utility, reproducibility, and impact beyond diagnosis and management of cancer.

References

1. International Agency for Research on Cancer, *Global Cancer Observatory: Cancer Today*. Lyon, France: IARC, 2024. [Online]. Available: <https://gco.iarc.who.int/today/>
HYPERLINK "https://gco.iarc.who.int/today/"
HYPERLINK "https://gco.iarc.who.int/today/"
HYPERLINK "https://gco.iarc.who.int/today/"
HYPERLINK "https://gco.iarc.who.int/today/"
HYPERLINK "https://gco.iarc.who.int/today/"
HYPERLINK "https://gco.iarc.who.int/today/"
HYPERLINK "https://gco.iarc.who.int/today/"
HYPERLINK "https://gco.iarc.who.int/today/"
HYPERLINK "https://gco.iarc.who.int/today/"
HYPERLINK "https://gco.iarc.who.int/today/"
2. World Health Organization, "Promoting cancer early diagnosis," Geneva, Switzerland: World Health Organization, 2024. [Online]. Available: <https://www.who.int/activities/promoting-cancer-early-diagnosis>
HYPERLINK "https://www.who.int/activities/promoting-cancer-early-diagnosis"
HYPERLINK "https://www.who.int/activities/promoting-cancer-early-diagnosis"
HYPERLINK "https://www.who.int/activities/promoting-cancer-early-diagnosis"
HYPERLINK "https://www.who.int/activities/promoting-cancer-early-diagnosis"
HYPERLINK "https://www.who.int/activities/promoting-cancer-early-diagnosis"
HYPERLINK "https://www.who.int/activities/promoting-cancer-early-diagnosis"
HYPERLINK "https://www.who.int/activities/promoting-cancer-early-diagnosis"
HYPERLINK "https://www.who.int/activities/promoting-cancer-early-diagnosis"
HYPERLINK "https://www.who.int/activities/promoting-cancer-early-diagnosis"
HYPERLINK "https://www.who.int/activities/promoting-cancer-early-diagnosis"
3. World Health Organization, *WHO Report on Cancer: Setting Priorities, Investing Wisely and Providing Care for All*. Geneva, Switzerland: World Health Organization, 2020.
4. M. Cristiano *et al.*, "Genome-wide cell-free DNA fragmentation in patients with cancer," *Nature*, vol. 570, pp. 385–389, Jun. 2019.
5. M. Heitzer, A. Haque, C. T. Roberts, and I. G. Speicher, "Current and future perspectives of liquid biopsies in genomics-driven oncology," *Nature Reviews Genetics*, vol. 20, pp. 71–88, Feb. 2019.
6. J. Hasin, A. Seldin, and A. Lusis, "Multi-omics approaches to disease," *Genome Biology*, vol. 18, Art. no. 83, May 2017.
7. K. Swanson *et al.*, "From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment," *Cell*, vol. 186, no. 8, pp. 1772–1791, 2023.
8. M. A. Manzoor *et al.*, "Multimodality representation learning: A survey on evolution, pretraining and its applications," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 20, no. 3, pp. 1–34, 2023.
9. A. A. Seyhan, "Circulating liquid biopsy biomarkers in glioblastoma: Advances and challenges," *Int. J. Mol. Sci.*, vol. 25, no. 14, Art. no. 7974, 2024.

10. W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373–63394, 2019.
11. C. Alix-Panabières, "Perspective: The future of liquid biopsy," *Nature*, vol. 579, no. 7800, p. S9, 2020.
12. J. C. Laguna *et al.*, "Tackling non-small cell lung cancer in young adults: From risk factors and genetic susceptibility to lung cancer profile and outcomes," *Amer. Soc. Clin. Oncol. Educ. Book*, vol. 44, no. 3, Art. no. e432488, 2024.
13. M. Esteller, "Cancer epigenomics: DNA methylomes and histone-modification maps," *Nature Rev. Genet.*, vol. 8, no. 4, pp. 286–298, 2007.
14. M. J. Beyeler *et al.*, "Comparing tangible retinal image characteristics with deep learning features reveals their complementarity for gene association and disease prediction," *medRxiv*, 2024.
15. T. Moser *et al.*, "Bridging biological cfDNA features and machine learning approaches," *Trends Genet.*, vol. 39, no. 4, pp. 285–307, 2023.
16. L. Keller *et al.*, "Clinical relevance of blood-based ctDNA analysis: Mutation detection and beyond," *Br. J. Cancer*, vol. 124, no. 2, pp. 345–358, 2021.
17. F. Zhang *et al.*, "Pan-precancer and cancer DNA methylation profiles revealed significant tissue specificity of interrupted biological processes in tumorigenesis," *Epigenetics*, vol. 18, no. 1, Art. no. 2231222, 2023.
18. V. S. Patil, R. Zhou, and T. M. Rana, "Gene regulation by non-coding RNAs," *Crit. Rev. Biochem. Mol. Biol.*, vol. 49, no. 1, pp. 16–32, 2014.
19. M. Nikanjam, S. Kato, and R. Kurzrock, "Liquid biopsy: Current technology and clinical applications," *J. Hematol. Oncol.*, vol. 15, no. 1, Art. no. 131, 2022.
20. Y. Lei *et al.*, "Applications of single-cell sequencing in cancer research: Progress and perspectives," *J. Hematol. Oncol.*, vol. 14, no. 1, Art. no. 91, 2021.
21. E. Armingol *et al.*, "Deciphering cell–cell interactions and communication from gene expression," *Nature Rev. Genet.*, vol. 22, no. 2, pp. 71–88, 2021.
22. P. H. G. Sanches *et al.*, "Integrating molecular perspectives: Strategies for comprehensive multi-omics integrative data analysis and machine learning applications in transcriptomics, proteomics, and metabolomics," *Biology*, vol. 13, no. 11, Art. no. 848, 2024.
23. J. L. Sylman *et al.*, "The predictive value of inflammation-related peripheral blood measurements in cancer staging and prognosis," *Front. Oncol.*, vol. 8, Art. no. 347682, 2018.
24. D. Hussain *et al.*, "Revolutionizing tumor detection and classification in multimodality imaging based on deep learning approaches: Methods, applications and limitations," *J. X-Ray Sci. Technol.*, vol. 32, no. 4, pp. 857–911, 2024.
25. S. Steyaert *et al.*, "Multimodal data fusion for cancer biomarker discovery with deep learning," *Nature Mach. Intell.*, vol. 5, no. 4, pp. 351–362, 2023.
26. S. Chae *et al.*, "Predicting emergency department visits and hospitalizations for patients with heart failure in home healthcare using a time series risk model," *J. Amer. Med. Inform. Assoc.*, vol. 30, no. 10, pp. 1622–1633, 2023.
27. F.-Z. Nakach, A. Idri, and E. Goceri, "A comprehensive investigation of multimodal deep learning fusion strategies for breast cancer classification," *Artif. Intell. Rev.*, vol. 57, no. 12, Art. no. 327, 2024.
28. K. M. Boehm *et al.*, "Harnessing multimodal data integration to advance precision oncology," *Nature Rev. Cancer*, vol. 22, no. 2, pp. 114–126, 2022.
29. A. Waqas *et al.*, "Multimodal data integration for oncology in the era of deep neural networks: A review," *Front. Artif. Intell.*, vol. 7, Art. no. 1408843, 2024.