The American Journal of Engineering and Technology
ISSN 2689-0984 Volume 07 - 2025

AI-Driven Liquid Biopsy Analysis for Early Cancer Detection: Integrating Circulating Tumor DNA, RNA, and Protein Biomarkers

Shakhnoza Khamidova
Independent Researcher, USA

Received: 02 Dec 2025 | Received Revised Version: 13 Dec 2025 | Accepted: 23 Dec 2025 | Published: 31 Dec 2025

Volume 07 Issue 12 2025 |

Abstract

Liquid biopsy has recently proved to be a minimally invasive method used to identify cancer-associated molecular information contained in biological fluid such as circulating tumor DNA (ctDNA), cell-free DNA (cfDNA), circulating RNA, microRNAs (miRNAs), and tumor-associated proteins. The early detection of cancer is difficult, however, because the signals from the tumor are often at low levels in a complex biological background. The computational tools offered by artificial intelligence (AI), such as machine learning and deep learning, enable the extraction of informative molecular features, the generation of biomarker signatures, and the integration of heterogeneous liquid-biopsy datasets. In this review, AI approaches for the early detection of cancer based on ctDNA/cfDNA, circulating RNA, miRNAs, and protein biomarkers will be examined. It explores the biological properties of these analytes, the pre-analytical and analytical process, AI-powered feature selection and classification techniques, and the creation of individual and multimodal biomarker signatures. Special emphasis is placed on multimodal and multi-omics approaches to the integration of genomic, epigenetic, transcriptional, and protein information. The success of AI-driven liquid biopsy in real-world applications, such as those in specific cancers and multi-cancer early detection, underscores its promise in the detection of intricate molecular patterns that might not be detectable by a single biomarker. However, biological heterogeneity, low tumor fraction, pre-analytical variation, assay noise, high-dimensional data, overfitting, population bias, and lack of external validation are still significant obstacles to clinical translation. Scaling up with increasingly larger and more diverse cohorts, standardized workflows, explainable AI, multimodal integration, and prospective longitudinal validation will be needed to support future progress. In summary, AI-based liquid biopsy is a promising approach that holds the potential to transform the diverse landscape of circulating biomarkers into molecular signatures for early cancer detection; however, it must be carefully validated and proven to be clinically reproducible.

Keywords: Artificial intelligence; liquid biopsy; Early cancer detection; circulating tumor DNA; Cell-free DNA; Circulating RNA; MicroRNA; Protein biomarkers, machine learning, deep learning, multi-omics, multimodal biomarker integration, and cancer screening.

© 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). AI-Driven Liquid Biopsy Analysis for Early Cancer Detection: Integrating Circulating Tumor DNA, RNA, and Protein Biomarkers. The American Journal of Engineering and Technology, 7(12), 230–249. Retrieved from https://theamericanjournals.com/index.php/tajet/article/view/8524

Introduction

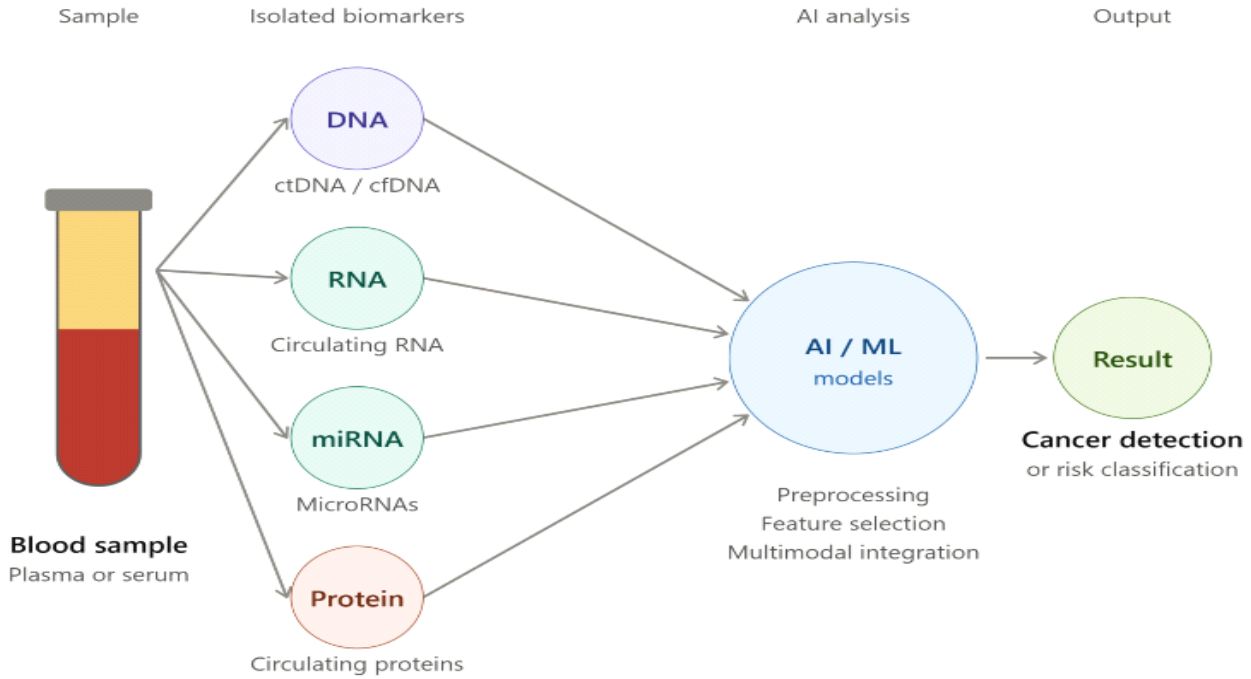
Cancer is still a leading cause of disease morbidity and mortality in the world, and the time of diagnosis of malignancy has a significant impact on treatment plan and prognosis. Early diagnosis will improve the chances that the disease is localized and can be cured, while late diagnosis can carry with it an increased burden of the disease, metastasis, and/or complex treatment needs. Clinically valuable screening methods such as mammography, colonoscopy, fecal testing, and low-dose computed tomography (LDCT) for specific high-risk groups have been established [1]. These strategies are, however, limited to certain cancers and certain populations, and in many cases, there are no standard methods available for the early detection of cancer.

Tissue biopsy is the backbone of cancer diagnosis, as it is the only way to directly obtain information on the morphology, histopathology, and molecular features of the tumor [2]. However, tissue sampling is an invasive procedure, can be technically challenging, depending on the tumor location, and is only a small area of a heterogeneous tumor. It may also not be feasible to take repeated tissue biopsies to monitor changes in molecules over time. These restrictions have fueled interest in minimally invasive means of acquiring molecular information from biological fluids associated with tumors.

One of the challenges in the blood-based detection of cancer is the relatively small quantity of material derived from the tumor, especially during the initial stages of the disease, when the tumor is small. The concentration of molecules derived from the circulating tumor may then be close to the analytical detection limits of available technologies. Therefore, early cancer detection may also depend on sensitive molecular assays as well as computational methods that can discriminate between the weak cancer-associated signal and a notable, biological, and technical background [3].

Liquid biopsy has become a new minimally invasive approach to the acquisition of molecular data from biofluids, such as blood. Liquid biopsy techniques are not an actual sampling of the tumor tissue but rather the analysis of entities that are released from the tumor or are a part of systemic reactions to malignancy. These encompass cell-free DNA (cfDNA), circulating tumor DNA (ctDNA), circulating RNA, microRNAs (miRNAs), extracellular vesicles, circulating tumor cells, and tumor-associated proteins. All these biomarkers can be used together to give complementary information about genetic changes, epigenetic status, transcriptional activity, and biological processes associated with tumors.

One of the most appealing features of blood is its ability to be captured multiple times, with relatively little procedure burden, allowing for a longitudinal evaluation of molecular changes. Biomarkers in circulation can also contain signals from various parts of the tumor and thus contain a wider spectrum of molecular information than a single tissue sample. Blood-based analysis is difficult, however, due to the presence of lots of material from normal tissues in the bloodstream. The variety and contents are also dependent on the type of cancer, the stage of the disease, the biological characteristics of the sample, the method of sample collection, and the method of analysis.



- Conceptual framework of AI-driven liquid biopsy for early cancer detection.

New high-throughput sequencing (HTS), digital molecular assays, and transcriptomic and proteomic technologies have produced ever larger and more complex liquid-biopsy

datasets [4]. These data can have high dimensions and/or nonlinear relationships that may not be easily captured by conventional statistical techniques. In the fields of drug discovery, artificial intelligence (AI), especially machine learning (ML), and deep learning (DL) can offer computational solutions for extracting complex molecular patterns, identifying informative features, and creating predictive models. These techniques have therefore become relevant in cancer research to obtain clinically relevant information from molecular datasets that are heterogeneous.

In liquid biopsy, ML algorithms can be used to detect patterns of molecular features that are linked to malignancy, while DL can be used to learn higher-level representations from genomic, transcriptomic, or proteomic data. The use of multiple biomarker classes is particularly important for the potential value of AI. All three, ctDNA, circRNA, and proteins, can give genomic, epigenetic, transcriptional, and regulatory information, and proteins can give information about cellular signaling and systemic biological responses [5]. These complementary molecular layers can provide patterns that are hard to identify with single biomarkers, but multimodal analysis has its own set of issues, including the need for normalization, feature selection, pattern interpretation, and validation.

The present review focuses on the use of AI tools for analysis of liquid biopsies in early cancer detection, especially on the analysis of ctDNA/cfDNA, circulating RNA, miRNAs, and tumor-associated proteins. It covers the biological properties of these biomarkers, feature extraction and representation at the molecular level, feature selection and classification using AI/ML techniques, development of biomarker signatures, and multimodal integration. Special focus is placed on the detection of weak signals from the tumor in a complex circulating background and the possible advantages of the combination of complementary classes of biomarkers. The review also addresses key challenges to clinical translation, such as biological heterogeneity, technical variability, small number of datasets, model generalizability, model interpretability, and prospective and independent validation. The overall goal is to assess the potential of AI to generate meaningful molecular signatures from heterogeneous liquid-biopsy measurements for early cancer detection and to define the methodological and clinical needs for successful implementation.

- **Biological Foundations of Liquid Biopsy Biomarkers**

Liquid biopsy allows the detection and molecular characterization of biological materials that are released into the circulation by tumor cells, the surrounding tissues, and host systems [6]. Circulating proteins, extracellular vesicle-associated molecules, circulating RNA, circulating tumor DNA (ctDNA), and cell-free DNA (cfDNA) are major analytes. The classes of biomarkers represent complementary aspects of tumor biology: DNA is used for genomic and epigenetic information, RNA for transcriptional and regulatory activity, and proteins for downstream cellular functions and tumor–host interactions. Complementary properties give rise to multimodal liquid-biopsy analysis, which is a biological basis.

- **Circulating Cell-Free DNA and Circulating Tumor DNA**
- **Origin and Biological Characteristics of cfDNA**

Cell-free DNA refers to short pieces of DNA that are floating in the blood, especially in plasma. It is released primarily because of apoptosis and necrosis but may be released by active cellular processes as well. Most cfDNA in healthy individuals is from non-malignant cells, as normal tissues continuously shed DNA. Cancer detection, then, depends upon detecting a small amount of tumor-derived cfDNA in a background of a much larger amount of normal cfDNA [7].

Tissue turnover, cell death, inflammation, physiological status, and disease status affect the concentrations and composition of cfDNA. Additional chromatin structuring and nuclease activity contribute to the characteristics of cfDNA fragments. Nucleosomes may be able to confer protection against degradation of DNA and may also give rise to characteristic fragment-size and genomic-distribution patterns. The presence of cfDNA, therefore, not only offers information via its sequence but also via the amount, the length of fragments, and the distribution within the genome.

- **Tumor-Derived DNA and ctDNA**

Circulating tumor DNA is the part of cfDNA that is from malignant cells or tumor tissue. The proportion of this in the total cfDNA is found to differ depending on the patient and the type of cancer and is typically low in patients with low tumor load and is referred to as the tumor fraction. This low abundance is a significant challenge to early detection because tumor molecules may be masked by cfDNA from normal cells [8].

Even though there is a small amount of ctDNA, it can carry some of the molecular features of the original tumor. The analysis can detect mutations, copy-number changes, methylation abnormalities, and other molecular changes associated with the tumor without the need to sample tissue. Therefore, cancer detection by ctDNA is one of the key molecular analytes studied in the context of blood-based cancer detection.

- **Genomic and Epigenetic Features of ctDNA**

There are several molecular characteristics of ctDNA that might be informative. Somatic mutations can send relatively specific signals for tumors, and copy-number alterations (CNAs) can give general patterns across the genome of the tumor that can be used for classification. DNA methylation can provide complementary information, as methylation patterns that are characteristic of cancer may be associated with tissue type, abnormal gene regulation, and malignant transformation.

The other important feature is the fragmentation of cfDNA. The chromatin structure, nucleosome positioning, transcriptional activity, and nuclease-mediated cleavage all affect the fragment size, genomic distribution, and end characteristics. These characteristics can then be helpful to infer information on tissues and cellular processes that shed DNA into circulation. The mutations, CNAs, methylation, and fragmentation can be combined to create a more comprehensive molecular signature than relying on a single feature, which can be especially useful in the context of machine-learning analysis of high-dimensional cfDNA data [9].

- **Circulating RNA Biomarkers**
- **Messenger RNA and Long Non-Coding RNA**

Circulating RNA is a molecular layer that is complementary to DNA. Messenger RNA (mRNA) is an indicator of gene expression, while long non-coding RNAs (lncRNAs) are involved in regulatory mechanisms such as transcription, chromatin structure, and cellular signaling. Changes in the RNA-expression pattern may thus reflect malignant transformation and changes in tumor-associated pathways.

Circulating RNA may give information relating to the functional state of cells compared with genomic DNA. RNA, however, is labile, and thus the abundance of RNA measured may be influenced by sample handling, RNA extraction, normalization, and the various cell types that give rise to the RNA in circulation. The factors mentioned should be carefully analyzed in biomarker development [10].

- **Circulating microRNAs**

MicroRNAs (miRNAs) are small, non-coding RNAs involved in posttranscriptional control of gene expression. Changes in miRNA expression have been linked to cancer-related processes such as proliferation, apoptosis, invasion, angiogenesis, and immune regulation. Their detectability and relative stability in circulation have rendered them valuable candidate liquid-biopsy markers.

Typically, studies of the ability of miRNAs to detect cancer look at multi-miRNA signatures instead of individual miRNAs. Machine-learning techniques are appropriate to the task, as they can analyze combinations of expression features and to detect patterns that are not obvious from single biomarkers. The miRNA profile may be different, however, depending on the type of cancer, the stage of the disease, the population, the type of sample, and the analysis platform, hindering its reproducibility and generalization.

- **Extracellular Vesicle-Associated RNA**

Extracellular vesicles (EVs), such as exosomes and other populations of vesicles, are membrane-enclosed entities that are released and circulate in the extracellular environment. They can carry mRNA, miRNA, other non-coding RNA, DNA, and proteins. They can shield molecular cargo from degradation, and their function in intercellular communication allows a means of carrying tumor-associated information.

Molecular features of tumor-derived EVs can contain features that mirror the tumor and interactions between the tumor and microenvironment. But measurements can vary significantly with EV heterogeneity, isolation methods, populations of vesicles, and analytical procedures. Standardized workflows are thus crucial for the development of reliable EV-based biomarkers.

- **Stability and Biological Significance of Circulating RNA**

Changes in transcription, pathways, tissue-specific activity, and tumor-host interactions may be reflected in circulating RNA and not directly in mutations in the DNA. RNA measurements, however, are affected by pre-analytical and analytical variation. To ensure reliable analysis, sample processing must be controlled, extraction should be proper, and the samples must be normalized and quality checked [11].

- **Circulating Protein Biomarkers**
- **Tumor-Associated Proteins**

Proteins give information regarding the consequences of variations in the genome and/or the transcription of the genes. Tumor-associated proteins can be excreted directly by cancer cells or generated by stromal, immune, and other cells of the host because of the growth of the tumor. Thus, measuring protein allows both tumor-associated processes and systemic responses to be measured.

- **Cytokines and Signaling Proteins**

Communication between tumors and the microenvironment can be reflected in cytokines, growth factors, and signaling proteins. As part of tumor growth, immune reactions, angiogenesis, and remodeling of tissues, levels of inflammatory and signaling proteins may change. But there are many proteins that are also affected by inflammation, infection, age, and metabolic disorders. So individual proteins may not be so specific, and multi-protein panels or combining with DNA and RNA biomarkers may be more suitable.

- **Proteomic Signatures for Cancer Detection**

Proteomic technologies enable the measurement of many proteins in the blood at once, yielding high-dimensional data for computational modeling. Combinations of protein concentrations can be detected through AI and statistical learning methods, which look for patterns that are linked to cancer, even when they don't involve specific markers. Proteins reflect downstream functional and phenotypic processes, and thus, their integration with ctDNA and RNA can offer complementary information for cancer classification.

- **Complementarity of DNA, RNA, and Protein Biomarkers**

The classes of the major liquid-biopsy biomarkers are different and have different biological information, as shown in Table 1. The primary information from ctDNA is genomic and epigenetic, circulating RNA is primarily involved in transcriptional and regulatory activities, and proteins represent downstream functional processes [12].

1. Major liquid-biopsy biomarker classes and their characteristics

Biomarker	Molecular information	Major analytical features	Advantages	Limitations
cfDNA/ctDNA	Genomic/epigenetic	Mutations, methylation, CNAs, fragmentation	Tumor-associated molecular information	Very low abundance in early cancer
miRNA	Post-transcriptional	Expression patterns/signatures	Relative stability and regulatory information	Biological heterogeneity
mRNA/lncRNA	Transcriptional	Expression/signatures	Functional molecular information	Degradation and normalization challenges
Proteins	Functional/phenotypic	Concentrations/proteomic signatures	Functional and clinically measurable information	Limited specificity of individual proteins
Multimodal biomarkers	Integrated	Combined molecular features	Complementary biological information	High dimensionality and integration complexity

The biological level of information has three components: DNA contains genetic and epigenetic information, RNA contains transcriptional and regulatory information, and proteins contain functional information at the cellular and systemic level. The combination of these layers could thus offer a more comprehensive picture of malignant processes than the analysis of one class of biomarkers.

This complementarity is especially pertinent when it comes to cancer detection through AI. Some low-level evidence in one category of biomarkers can be corroborated by another category. Examples include genomic changes that can infer the origin of the tumor, RNA signatures that can infer changes in cellular activity, and protein signatures that can infer changes in the tumor-host interactions. Multimodal AI models can train the relationships between these molecular layers instead of training each biomarker individually.

Multimodal integration introduces several other challenges, such as dimensionality, missing data, normalization, and model complexity, however. The successful integration of this data, however, depends on the selection of the features, data harmonization, interpretability, and independent validation. The next sections present the multimodal AI strategies, which are grounded in these biological and analytical components.

- **Pre-Analytical and Analytical Workflow of AI-Enabled Liquid Biopsy**

Computational models and the quality of biological samples given to AI-driven liquid-biopsy assays are critical to their performance. Steps in the workflow involve blood collection, preparation of plasma or serum, biomarker extraction, molecular profiling, preprocessing, feature representation, and classification using AI. Any change in the measurement of a biomarker in the circulation can affect the biomarker and, therefore, the performance of the model. This is especially critical in early cancer detection, where signals from a tumor may be very low. Standardized workflows are, therefore, key to reproducible molecular measurements and reliable computational predictions.

- **Blood Collection and Sample Processing**

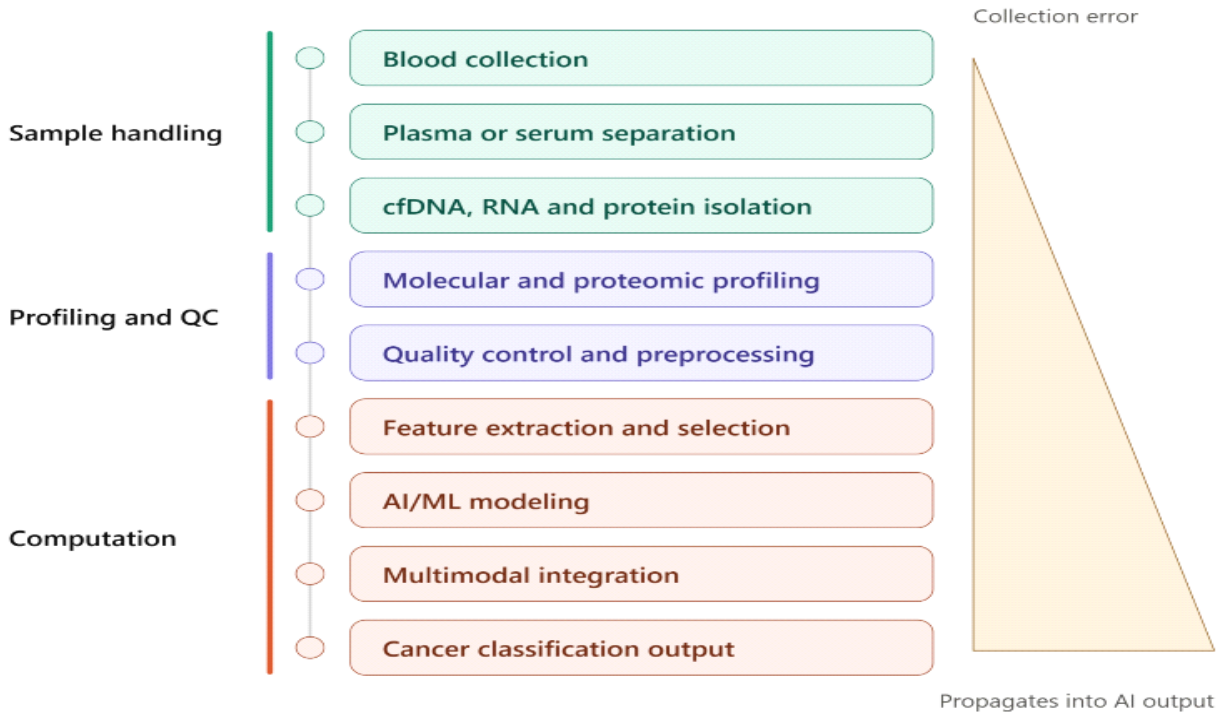
The contained DNA, RNA, proteins, and extracellular vesicles in the blood are accessible. However, cfDNA and ctDNA analysis is typically performed in plasma due to the presence of clotting in serum that could release genomic DNA from leukocytes, leading to background DNA levels. Biomarker integrity and abundance can be impacted by collection tubes and collection tube processing time, anticoagulants, centrifugation, transport, and storage conditions [13].

Repeated freeze–thaw can impact biomolecules, and delayed processing can lead to increased cell degradation and genomic contamination. AI is especially sensitive to these pre-analytical differences, as they can be learned as artificial disease-related patterns. Therefore, the handling procedures of the samples and the associated metadata should be standardized and documented.

- **Biomarker Extraction and Molecular Profiling**

Plasma or serum preparation yields isolation and measurement of biomarkers on molecular and analytical platforms. The next generation of sequencing using targeted sequencing, genome-scale sequencing, methylation assays, digital PCR, and fragmentomic analysis can be used to evaluate ctDNA/cfDNA. Different types of mutations, copy number changes, methylation profiles, and characteristics of the fragments can then be used as computational features.

To measure mRNA, miRNA, lncRNA, and other types of RNA, circRNA can be analyzed by sequencing or targeted assays. Normalization procedures and RNA degradation must be undertaken with care due to the sensitivity of RNA to degradation. Immunoassays, targeted assays, mass spectrometry, and multiplexed proteomic platforms can all be used to measure protein biomarkers. Multiple protein datasets can be combined to form a high-dimensional dataset that is well suited for AI, as multiple proteins can give more classification signals than a single protein alone.



End-to-end workflow from blood collection to AI-based cancer classification.

- **Data Preprocessing and Quality Control**

For you to use AI to analyze molecular data, it needs to be preprocessed. Read quality assessment, alignment, duplicate filtering, variant filtering, and sequencing depth can be performed on sequencing datasets. Typically, RNA and protein data sets need to be normalized, transformed, and have low-quality measurements removed and technical variation corrected.

These are critical tasks to perform due to potential artificial patterns caused by sequencing errors, assay variation, multiple processing batches, and incomplete measurements. AI models can learn lab-specific characteristics if correction is not performed. With consistent preprocessing and reporting, it is then easier to achieve reproducibility and model generalizability.

- **Feature Extraction and Representation**

Once the samples are quality controlled, molecular measurements are assembled to be suitable for use in an AI model. The mutations, copy number alterations, methylation, fragment lengths, and end pattern may be present in ctDNA. The abundance of transcripts and miRNAs is represented as RNA features, and concentrations, ratios, and protein patterns are represented as protein features [14].

Feature selection and dimensionality reduction are relevant since liquid-biopsy datasets can be extremely large, with thousands of variables but a few signals related to the disease. Biologically meaningful information can be captured while retaining computational efficiency and avoiding overfitting the data by using appropriate representation. Therefore, establishing reliable sample preparation, standardizing sample processing, and providing meaningful and informative feature representation are essential for the accurate classification of liquid biopsies using AI.

- **Artificial Intelligence Methods for Liquid Biopsy Analysis**

Numerous DNA, RNA, protein, and clinical variables can be included in liquid-biopsy data sets, with many having weak or nonlinear associations with cancer. Methods for extracting informative patterns from these high-dimensional datasets are available through the lens of artificial intelligence (AI), such as those from traditional machine learning (ML) and deep learning (DL). They are effective based on the size of the samples, the structure of

the features, the quality of the data, and the clinical application they are intended for.

- **Conventional Machine Learning Approaches**
- **Logistic Regression**

Logistic regression is a tool widely used to make binary classifications and can also provide estimates of the probability that a sample is an "outlier" to a cancer or control group. Parameters that can be used as predictors in a liquid biopsy include mutation frequencies, methylation status, RNA expression, and concentration of proteins. Its greatest strength is that it is relatively interpretable, meaning that the contribution of individual variables to the model can be seen in the form of model coefficients. Regular logistic regression, however, is not very well suited to representing complex nonlinear interactions [15]. Therefore, L1- and L2-regularized forms are well suited for high-dimensional datasets of biomarkers, for controlling the complexity of these models, and, in the case of L1 regularization, for feature selection.

- **Support Vector Machines**

Support vector machines (SVMs) are used to determine the decision boundaries that maximize the separation between classes. SVMs can be used to represent nonlinearities and are therefore applicable to molecular data sets that may not be linearly separable. They can function with many molecular features and limited numbers of samples. The performance, however, requires the selection of the kernel, regularization, and optimization of the parameters. Therefore, it is important to have independent validation to ensure generalizability.

- **Random Forest and Ensemble Learning**

Random forests are a collection of decision trees that were used to model the nonlinear relationship and interaction between molecular variables. They can also give measures of variable importance, aiding in the identification of biomarkers. The case of several weakly informative biomarkers, which together can provide more discrimination, is relevant for ensemble learning. But correlated features may share their importance among features, and complex ensemble models can be less interpretable than simple models.

- **Gradient-Boosting Methods**

Gradient-boosting algorithms build decision trees one by one to add to the errors of the previous trees to enhance the predictions. Non-linear relationships and interactions between molecular and clinical features can be captured via methods like XGBoost. They are helpful in merging heteroscedastic liquid biopsies measurements but can overfit high-dimensional data sets with small numbers of samples. Hence, the need for cross-validation, regularization, feature selection, and independent testing [16].

2. AI/ML methods applicable to liquid-biopsy analysis

AI approach	Typical application	Strength	Limitation
Logistic regression	Biomarker classification	Relatively interpretable	Limited nonlinear modeling
SVM	High-dimensional classification	Effective with smaller datasets	Parameter/kernel sensitivity
Random forest	Biomarker selection/classification	Captures nonlinear relationships	Reduced interpretability
Gradient boosting	Risk/classification modeling	Strong nonlinear capability	Overfitting risk
Neural networks	Complex/multimodal analysis	Learns nonlinear representations	Greater data requirements

Autoencoders	Feature representation	Dimensionality reduction	Interpretation challenges

- **Deep Learning Approaches**
- **Neural Networks**

Artificial neural networks (ANN) learn nonlinear relations by using interconnected layers of computational units. They can handle the extraction of ctDNA, RNA, protein, or several molecular characteristics and can detect multiple weak biomarker interactions. However, neural networks in general need sufficient training data, proper regularization and validation, and sometimes, abstract representations within the network are not easily interpretable [17].

- **Convolutional and Sequence-Based Models**

Structured genomic or molecular representations can be used to train a convolutional neural network (CNN) to learn about local patterns of related features. In recurrent and other models based on sequences, ordered information like a sequence of nucleotide bases or a sequence of measurements of a biomarker can be analyzed. When there is a meaningful sequence or temporal structure, their application is promising, but it requires enough data and data that are collected consistently.

- **Autoencoders**

Autoencoders are deep learning models that can learn low-dimensional representations of high-dimensional data by mapping the data into a lower-dimensional space and then mapping it back to its original form. They can help to decrease the dimensionality and aid in the representation of genomic, transcriptomic, or proteomic profiles in liquid biopsy. For latent features, however, it may be more difficult to directly correlate with recognizable biological biomarkers [18].

- **Feature Selection and Dimensionality Reduction**

High-dimensional liquid-biopsy datasets create a substantial risk of overfitting. Feature selection therefore aims to identify informative variables while reducing unnecessary complexity.

Features can be selected using statistical methods based on differential expressions, association tests, or correlations. But weak features can be informative when combined. LASSO regression overcomes this issue by introducing an L1 penalty, which may lead to a shrinkage to zero of some of the coefficients, while simultaneously providing prediction and feature selection. Correlated biomarkers will also have to be interpreted carefully, as LASSO might keep just one out of a set of related features.

Principal component analysis (PCA) is a technique that reduces a set of correlated variables to a smaller set of variables that account for the large sources of variation. PCA can help to visualize and reduce dimensionality, but the resulting components are not necessarily biologically interpretable [19].

Variables are selected using embedded and wrapper methods such as tree-based selection and recursive feature elimination, based on the contribution of the variables to the model performance. It is important to note that feature selection should only take place in the training and validation fold to avoid information leakage from test data.

- **AI for Biomarker Signature Development**

Multi-biomarker signatures, instead of individual ones, are more the focus of AI-based analysis. Due to the varying abundance of biomarkers in patients and disease stage, a single mutation, miRNA, or protein can be limited as a diagnostic protein with low sensitivity or specificity. One can consider the combination of several features that may capture complementary aspects of tumor biology [20].

Signatures can encompass mutations, methylation, copy number, fragmentomic features, RNA expression, protein levels, and clinical parameters. Models can be developed for cancer type or multi-cancer classification. The more biomarkers, however, the more complicated, expensive, and prone to overfitting the assay becomes. A good signature should therefore be a compromise between predictive power, reproducibility, biological relevance, and measurement practicality.

- **Model Interpretability and Explainable AI**

The importance of interpretation in understanding which molecular variables contribute to AI predictions. Researchers may use feature-importance methods and explainable AI (XAI) approaches to explore the influence of individual biomarkers. At the individual level, for instance, SHapley Additive exPlanations (SHAP) allow estimating the impact of single features on the model prediction, which can be extended at the population level.

In liquid-biopsy research, XAI can uncover patterns of genomic, RNA, and protein characteristics that aid in the diagnosis of cancer and can help assess biological plausibility. However, prediction is not causation. A feature closely correlated with a model prediction can be for a variety of reasons, including technical variation, correlation, or another indirect association that does not provide any mechanism for cancer.

In general, AI techniques can span from simple statistical models, which are easy to understand, to more sophisticated neural architectures that can learn from what they are trained on, such as nonlinear and multimodal representations. Appropriate algorithm selection, rigorous feature selection, robust validation, and meaningful interpretation are crucial for their successful application in liquid biopsy. It is not just about accuracy of cancer classification; it's about building reproducible and biologically meaningful molecular signatures to detect subtle cancer-related signals during complex circulating markers.

- **AI Analysis of Individual Liquid-Biopsy Biomarker Classes**

The analytes obtained from liquid biopsy can offer a variety of molecular information, which can be analyzed with AI. The advantage of ctDNA/cfDNA is a combination of genomic, epigenetic, and fragmentomic information, whereas circulating RNA contains information about transcriptional and regulatory activity, and proteins contain information about function and host response. AI can detect patterns of molecular characteristics that may be linked to cancer, as individual markers are not sufficient to make a diagnosis.

- **AI-Based ctDNA and cfDNA Analysis**
- **Mutation-Based Detection**

In mutation-based ctDNA analysis, tumor-specific mutations are detected in a patient's bloodstream. Early-stage cancers may shed just a few ctDNA molecules, however, and thus individual mutations may not be detectable. AI can analyze the frequency of variants, multiple mutations, and clinical characteristics to uncover composite cancer patterns. This can be useful in cases where one mutation doesn't have that much prevalence but with low tumor fraction and limited informative mutations [21].

- **Methylation-Based Detection**

DNA methylation changes in cancer cells carry information other than genetic changes and can include signals of tissue of origin [27]. Methylation datasets can span thousands of genomic positions, and machine-learning algorithms can detect patterns of informative positions and build a signature of cancer. However, methylation patterns are tissue- and/or individual-specific, and normal cfDNA may mask the presence of tumor-specific methylation signals. There are also differences between the assay platforms and sample processing that must be carefully standardized and validated independently.

- **Fragmentomics and Copy Number Analysis.**

Fragmentomics involves studying DNA fragment size, genomic distribution, end patterns, and features associated with the nucleosome. The properties of these may give insight into the cellular origin and biological mechanisms generating cfDNA. AI is particularly well suited to the analysis of the many potential fragmentomic features and finding patterns that can be hard to see with individual thresholds.

Another source of genomic information is copy-number alterations. Amplifications and deletions may help identify more general trends of genomic instability and can be used in conjunction with mutation, methylation, and fragmentation features. These methods enable AI models to analyze several aspects of circulating DNA, instead of just one molecular feature [28].

- **AI-Based Circulating RNA Analysis**
- **Signatures of miRNA, mRNA, and lncRNA**

The regulatory functions and the relative stability of the circulating miRNAs make them the subject of extensive research. But the specificity of miRNAs to specific cancers is not always high. AI can thus find multi-miRNA signatures that can discriminate between cancer and non-cancer samples or between the types of cancer [22].

Other information on transcriptional and regulatory changes is acquired by the circulating of mRNA or lncRNA. Machine-learning models are able to process many expression data sets and determine sets of expressions associated with distinct cancer phenotypes. However, there are issues with RNA degradation, normalization, tissue contribution, and platform differences that may impact the results' repeatability.

- **RNA-Based Cancer Classification**

AI models using RNAs can aid in the diagnosis of cancer, the differentiation of cancer from benign conditions, and the classification of cancer types. They can assess inter-transcript interactions and can therefore provide a moderately informative RNA as a part of a diagnostic signature. But the larger the expression data sets, the greater the risk of overfitting, and therefore feature selection and independent validation are required.

- **AI-Based Protein Biomarker Analysis**
- **Protein Panels and Proteomic Pattern Recognition**

AI allows for the analysis of multiple proteins in circulation at the same time instead of individual proteins. Protein panels can be used to detect signals of the tumor, inflammation, and host response. For instance, the DE cancer framework has shown the potential of machine learning to identify more compact sets of proteins that still provide valuable cancer classification information.

There can be hundreds or thousands of variables in a high-throughput proteomic experiment. Random forests, SVMs, gradient boosting, regression, and neural networks can be used to find combinations of proteins that are related to cancer. Inflammation and other non-malignant diseases can also affect protein levels;however, models need to be able to separate cancer-associated changes from the broader physiological response patterns.

- **Protein-Based Cancer Classification**

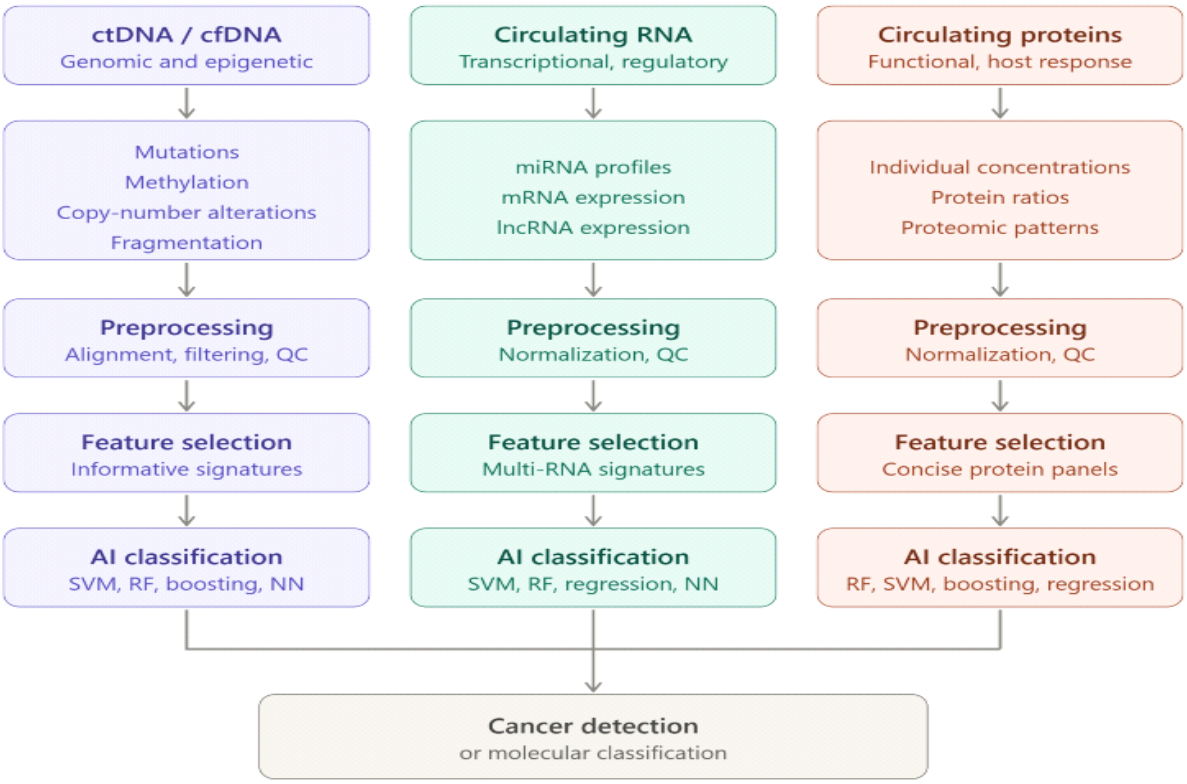
Protein-based classifiers can separate cancer from control patients and can also help to differentiate the type of cancer. The usefulness of their practical applications will rely on the reproducibility and dimensions of the protein panel. Very large panels can be used to enhance research-level classification, but smaller panels can be more easily standardized and made into routine diagnostic assays [23].

- **Comparative Characteristics of AI-Analyzed Biomarkers**

The three classes of biomarkers are not redundant. Genomic and epigenetic abnormalities can be detected by DNA, gene-regulatory activity can be detected by RNA, and downstream functional and host responses can be detected by proteins. These measurements are heterogeneous, and while they can be converted by AI into molecular signatures, each class has its own unique technical and biological limitations.

3. Comparison of AI applications across liquid-biopsy biomarker classes

Biomarker	Input/features	Common AI approaches	Advantages	Key limitations
ctDNA/cfDNA	Mutations, methylation, CNAs, fragments	SVM, RF, boosting, neural networks	Direct genomic/epigenetic information	Low tumor fraction, technical noise
miRNA	Expression and combinations	SVM, RF, regression, neural networks	Relative stability, regulatory information	Heterogeneity, limited specificity
mRNA/lncRNA	Transcript expression	Regression, SVM, RF, neural networks	Functional/transcriptional information	Degradation, normalization
Proteins	Concentrations, proteomic patterns	RF, SVM, boosting, regression	Functional and host-response information	Limited specificity of individual proteins
Multimodal	DNA + RNA + protein	Ensemble and multimodal ML/DL	Complementary biological information	High dimensionality and integration complexity



AI analytical strategies for individual liquid-biopsy biomarker classes.

Figure 3 shows three parallel pathways: ctDNA/cfDNA with mutation, methylation, copy-number, and fragmentation features; circulating RNA with miRNA, mRNA, and lncRNA profiles; and proteins with individual concentrations and proteomic patterns. Each pathway should proceed through preprocessing, feature selection, and AI classification toward cancer detection or molecular classification.

Importantly, early cancer signals can be seen spread over several aspects of molecular dimensions. There may also be a low level of ctDNA with altered methylation or fragmentation, or RNA and protein changes can also give clues to the presence of tumor activity. This explains the biological and computational justification for combining

DNA, RNA, and protein measurements in multimodal AI models as discussed.

- **Multimodal AI Integration of DNA, RNA, and Protein Biomarkers**
- ***Rationale for Multimarker Integration***

Liquid-biopsy biomarkers give individual information about tumor biology. ctDNA/cfDNA can be used to detect mutations, copy-number variations, methylation, and fragmentation, all of which are associated with transcriptional and regulatory activity, and proteins can be used to detect functional information and host response. Low signaling intensity and complexity of early tumors suggests that no single class of biomarkers may be representative.

Multimarker integration is designed to fuse these loosely coupled markers and minimize the reliance on a single weak and obscure marker. Genomic changes can, for instance, indicate the presence of tumor-derived material, whereas methylation, RNA, and protein characteristics can add information regarding the type of tissue, its activity, and the host response. The multi-omics approach is therefore able to produce molecular signatures that are coordinated to each other and may not be attainable from one analyte.

However, the use of combinations of modalities does not necessarily equate to better diagnostic performance. The dimensionality of the data is raised, and integration may lead to missing data, data platform-specific differences, normalization issues, and overfitting. Stages of careful feature selection and validation are therefore crucial [24].

- **Strategies for Multimodal Data Integration**
- **Early Data Fusion**

Early fusion: This fuses multiple features across various biomarker classes together in one feature matrix prior to model training. Simultaneous analysis of mutation, methylation, fragmentomic, RNA-expression, and protein features is thus possible. This method is useful to have a molecular representation that is consistent across all different modalities, and there is a need for careful modality normalization and feature selection due to varying scales, dimensionality, and noise in modalities.

- **Intermediate Fusion**

The first step of the intermediate fusion is to map each modality to a lower-dimensional space and then fuse them in a joint model. Latent features specific to modality (DNA, RNA, or protein) and latent features common to all modalities can be learned from the corresponding datasets using neural networks and autoencoders. This can be used to model cross-modal relations but is typically much harder to compute, needs larger sample sizes, and requires higher levels of regularization.

- **Late Fusion**

Late fusion forms separate models for each biomarker class and fuses the predictions from these models by weighted averaging, meta-classifier, or other ensemble techniques. This approach is flexible and could be adaptable if some data were not available. Some of the direct relationships between the molecular modalities will be lost, however, since interaction is primarily at the prediction stage.

4. Representative strategies for multimodal liquid-biopsy data integration

Integration strategy	Description	Potential benefit	Major challenge
Early fusion	Combines features before model training	Unified representation	Dimensionality and feature imbalance
Intermediate fusion	Learns modality-specific/shared representations	Captures cross-modal relationships	Model complexity and data requirements
Late fusion	Combines independent model predictions	Flexible and modular	May miss cross-modal interactions
Multi-omics learning	Joint analysis of multiple molecular layers	Greater biological comprehensiveness	Harmonization and missing data

- **Multi-Omics Machine Learning**
- **Genomic–Transcriptomic Integration**

Genomic and transcriptomic data provide various levels of tumor biology. DNA alterations provide a molecular profile of abnormalities, and RNA profiles provide a transcriptional and regulatory profile. Mutations, copy number variation, methylation, miRNAs, mRNAs, and lncRNAs can therefore be used as a more comprehensive molecular representation. May be useful in situations where genomic signals are weak, but important sources of variation include RNA degradation, normalization, and cellular composition.

- **Genomic–Proteomic Integration**

Genomic features are molecular changes, while proteins are functional changes that reflect downstream biological activity and host responses. These layers can be integrated to link genomic signals associated with tumors with their functional phenotypic information. While protein measurements are also subject to the influence of non-malignant conditions, AI can leverage these

complementary measurements to make up for when one modality is an incomplete signal.

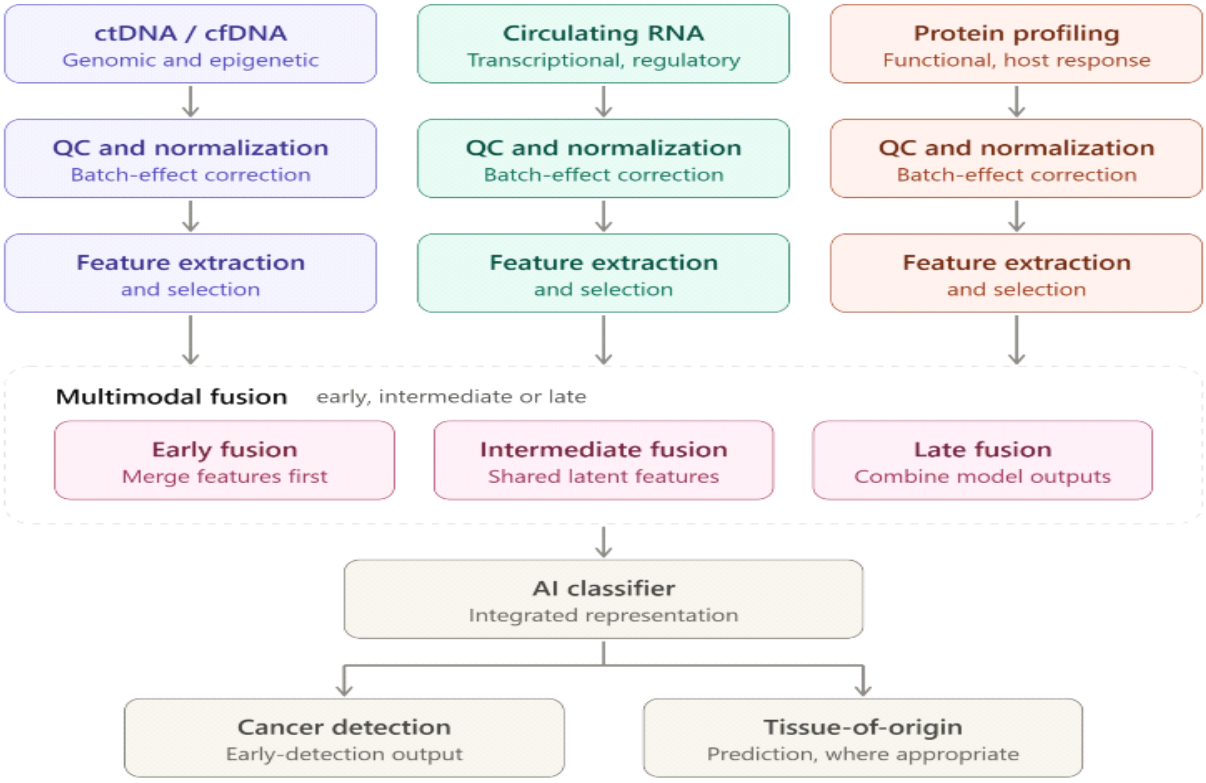
- **DNA–RNA–Protein Integration**

Combining DNA, RNA, and protein measures can give a more complete picture of the biology of malignancy and help tie together the genomic changes, the epigenetic changes, and the changes in transcription with the downstream effect of proteins. There is some evidence from multimodal studies that this approach is of value. A study published in 2024 found that their methylation method, which included the 5' end motifs, cfDNA methylation, copy-number variation, and mutations, compared to an independent validation cohort, had an AUC of 0.981 with 92.1% sensitivity and 94.7% specificity. Coordinated by the Intermountain Laboratory, another study on ovarian cancer in 2024 found that combining serum miRNA, protein markers, and clinical data provides better classification.

- **Development of Integrated Molecular Signatures**

When dealing with multimodal signatures, feature selection must be done with care to eliminate "noise," redundant, unstable, and technically driven signature variables. Regularization, dimensionality reduction, recursive feature elimination, and embedded selection methods can facilitate identification of informative combinations. Appropriate features weighting and generalizability, should be taken into account when optimizing the model, as the contribution of the molecular modalities might not be equal [25].

Cross-validation is useful to get an idea of the stability of the model, but external validation is still necessary. To avoid information leakage and overly optimistic estimates of performance, feature selection, normalization, and dimensionality reduction are required to be performed also within training folds.



Multimodal AI architecture for integrating ctDNA, RNA, and protein biomarkers.

Figure 4 shows three parallel streams for ctDNA/cfDNA, circulating RNA, and protein profiling. Each stream undergoes quality control, normalization, feature extraction, and selection before convergence through early, intermediate, or late fusion. The integrated representation is then processed by an AI classifier to generate cancer detection and, where appropriate, tissue-of-origin predictions.

Overall, multimodal AI offers a blueprint of how to integrate complementary molecular signals to enhance characterization of early malignancy. It is not only model

architecture that makes for its success, but also standardized data generation, an appropriate integration strategy, strict validation, and dimensionality and overfitting control.

- **Applications in Early Detection of Cancer**
- **Early Detection of Specific Cancer Types**

The performance of the AI-assisted liquid biopsy has been explored in a range of cancer types and is affected by cancer type, stage, biomarker level in the tumor, study design, etc.

cfDNA methylation, mutations, fragmentomics, and protein characteristics have been well studied to identify lung cancer. A combination of cfDNA methylation with protein biomarkers (multimodal approaches) is designed to encompass both tumor-associated molecular changes and systemic responses. Additionally, AI has been investigated in conjunction with imaging for the assessment of pulmonary nodules.

Colorectal cancer is an important application since molecular screening could help with detection prior to clinical symptoms. cfDNA methylation and 5′ end motifs, copy-number variation, and mutations have been incorporated into AI models. The results of a 2024 study showed an AUC of 0.981, 92.1% sensitivity, and 94.7% specificity in their validation cohort, a testament to the potential of complementary features of cfDNA. In addition, several studies have examined circulating miRNAs, but there are conflicting reports.

Circulating miRNAs, cfDNA methylation, genomic alterations, and proteins are all targets of breast cancer research. However, the specificity of individual miRNAs is often not high enough, and therefore, AI-based multi-miRNA signatures or multi-marker models have been explored. Epigenetic instability analysis using machine learning of cfDNA methylation was also reported for the early detection of breast and lung cancer.

The mutations, methylation, fragmentomics, circulating RNA, and proteins of liver cancer have been explored. The integration of genomic and epigenomic data could help in the discrimination in cases where signals from the tumor are few. AI feature selection tools can detect the combinations of features in molecular information that are informative while discarding the rest.

Since the signs of pancreatic cancer can be vague, it is a significant early-detection problem. Therefore, circulating DNA, RNA, and proteins have been studied. The OncoSeek study utilized seven protein tumor markers, each of which was classified with AI, in nine cancer types and found a 51.7% overall sensitivity and 92.9% overall specificity, with the highest sensitivity of 77.6% for pancreatic cancer. These results are representative of what a protein panel can achieve and show the variability for performance across the different types of cancer.

Other cancers, such as ovarian, gastric, esophageal, and head-and-neck cancers, have also been shown to be a potential target for AI-driven liquid-biopsy strategies. In 2024, a clinical report of stage-I ovarian, renal, and oropharyngeal cancers detected with a machine-learning-based test that analyzes cfDNA methylation was published. One such instance is proof of feasibility, not of effectiveness of screening in a population.

- **Multi-Cancer Early Detection**

Multi-cancer early detection (MCED) seeks to detect cancer information from numerous cancers with a single blood test. One of the most beneficial aspects of AI is that these methods can be based on thousands of features, such as methylation, the genome, fragments, RNA, or protein.

One of the prominent approaches in MCED is methylation-based cfDNA assays. A targeted methylation platform was analytically validated under controlled conditions with 99.3% specificity. A methylation approach based on the CCGA subsequently reported has an overall cancer sensitivity of 51.5% and a high specificity, while the sensitivity varied according to cancer type and stage. This was complemented by prediction of the probable tissue of origin [26].

Protein-based MCED is an alternative approach. In its test set, the DEcancer machine-learning framework achieved a higher detection sensitivity for eight cancers, ranging from 48% to 90%, and demonstrated that detection performance could be high even if the protein panel size was reduced.

- **Diagnostic Performance of AI-Based Models**

Sensitivity is the number of all cases of cancer that are identified correctly; it is especially important if cancer is detected early, since signals may be low from the tumors.

Specificity is the percentage of non-cancer patients that are rightly classified as negative. Ideally, population screening tests should be highly specific, as high numbers of false-positive tests can lead to a significant downstream diagnostic process.

AUC/ROC gives an overview of the discrimination at all classification levels. A good AUC value represents good

separation between two classes but is not necessarily a good AUC value at a clinical threshold relevant for screening.

The prevalence of diseases has a significant effect on both PPV and NPV. So, findings of enriched case-control studies cannot be simply applied to screening populations without symptoms.

A representative sampling of AI-powered liquid biopsy studies is outlined below in Table 5. Cancer stage, control population(s), biomarker platform, endpoints, and validation methodology vary and should not be directly compared across studies.

Overall, AI-enabled liquid biopsy is a blueprint for accessing weak molecular signals and combining complementary cfDNA, RNA, and protein data. It needs to be carefully validated in representative populations, with particular focus on the sensitivity and specificity in early diagnosis, predictive values, tissue-of-origin accuracy, positivity, and negativity.

- **Challenges, Limitations, and Clinical Translation**

While advances have been made in the use of AI to improve the liquid biopsy, there remain biological, technological, computational, and clinical challenges to translation into early cancer detection in the clinical setting. Cohort performance in high performance does not necessarily reflect performance in asymptomatic cohorts.

- **Biological Challenges**

One of the major challenges is the scarcity of tumor-derived signals. Early disease comes with the caveat that the amount of ctDNA is only a small portion of cfDNA; additional signals from normal tissues may also dilute the amount of circulating RNA and proteins. This means that meaningful molecular characteristics may also be considered as biological and technical noise and could therefore only be detectable based on the sensitivity of the assays and the AI classification.

Another issue of detection is the tumor heterogeneity, as different parts of a tumor may produce different molecular signals. Blood, however, only reflects systemic, but incomplete, tumor biology. Further, DNA, RNA, and protein profiles can be influenced by various factors, including age, sex, inflammation, comorbidities, medications, lifestyle, and tissue turnover. As a result, models developed from small populations can potentially learn patterns that are specific to that population, instead of patterns that are specific to cancer.

- **Technical and Analytical Challenges**

Pre-analytical variability associated with blood-collection tubes, processing time and rotation, centrifugation, storage period, and freeze–thaw cycles can greatly impact biomarker measurements. Processing delay can lead to higher background genomic DNA; RNA and protein measurements are impacted by degradation and handling. These differences could be mistakenly acquired as disease-related characteristics.

Weak tumor signals can also be masked by sequencing and assay noise, such as amplification bias, sequencing errors, poor sequencing coverage, or extraction efficiencies. For multimodal studies, there are additional risks due to batch effects among different laboratories, instruments, reagent lots, and assay platforms.

Another common scenario is that of liquid-biopsy datasets, which are also high-dimensional but still moderately sized, leading to a significant risk of overfitting. Therefore, feature selection, regularization, dimensionality reduction, and strict separation between the training and the test sets are essential.

- **AI-Specific Challenges**

The lack of early-stage cancer samples and lack of cancer type and stage representation can create limited and imbalanced datasets. Generalizability is also a significant worry, as models created in selected case control populations may vary in performance from institution to institution, region to region, between demographic groups, and between clinical settings.

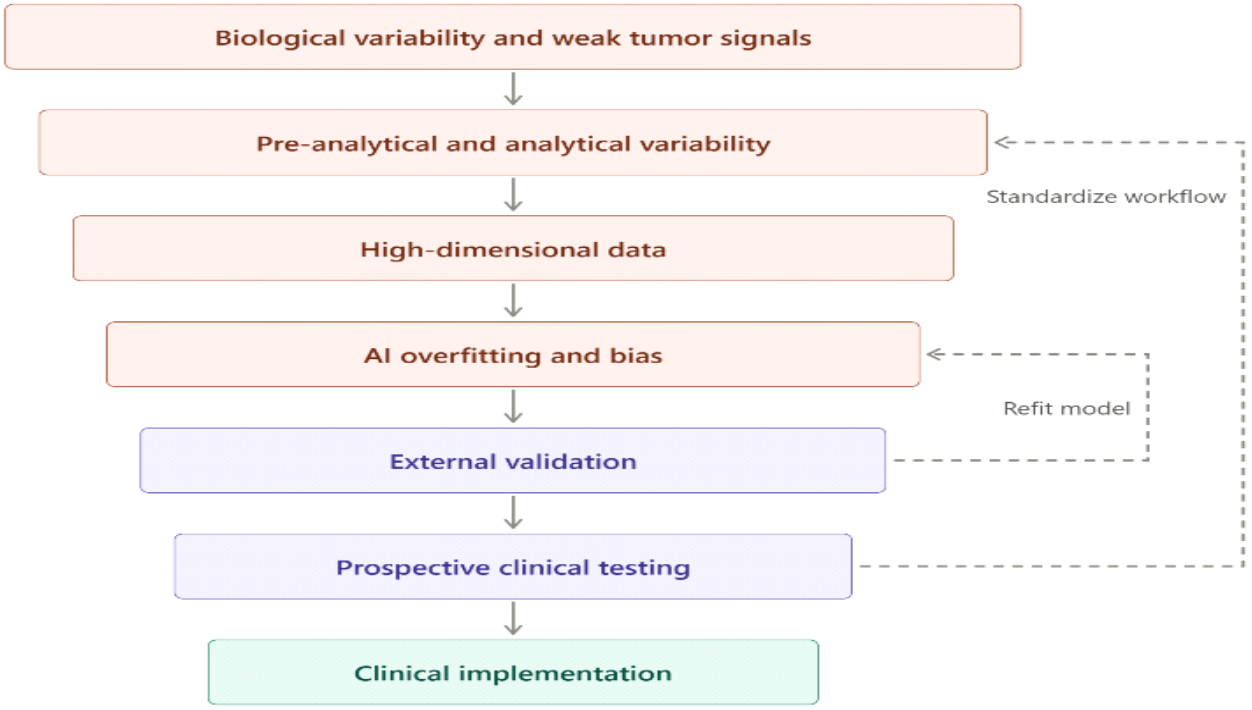
For complex ensemble and deep-learning models, interpretability is very important. Feature importance and SHAP can highlight variables that have a significant impact, but this does not imply that they are biologically causative. A population bias also needs to be assessed using a variety of cohorts and subgroups.

- **Validation and Clinical Translation**

There is an internal cross-validation that can estimate performance based on a development set but cannot assess transportability. Therefore, independent, multicenter external validation is needed. Cancer stages should be represented and realistic; non-cancer controls should be used, not unusual healthy volunteers.

Early detection, false positive/false negative implications, downstream diagnostic efforts, and overdiagnosis are all issues that need to be addressed in prospective clinical

validation. Standardization is also needed in the processing of samples, molecular assays, quality-control protocols, computational pipelines, and reporting procedures for clinical implementation.



Major barriers to clinical translation of AI-driven liquid-biopsy diagnostics.

- **Future Directions and Research Priorities**

Development of AI liquid biopsy will need advances in algorithmic performance as well as biological characterization, lab standardization, validation, and clinical implementation. The literature identified several research priorities that will help move forward early-cancer detection.

- **Development of Larger and More Diverse Cohorts**

Larger, multicenter populations across various ages, genders, geographic populations, cancer types, cancer stages, and comorbidities are necessary for future studies. Stage I and preclinical disease should be given a higher priority since models built mostly to study advanced cancers may overestimate the performance of early detection. Diverse cohorts can also help facilitate evaluation of the robustness of the model and population-specific biases.

- **Standardization of Liquid-Biopsy Workflows**

The standardization of blood collection, processing, storage, extraction of biomarkers, sequencing, RNA measurement, proteomic assays, and quality control procedures are essential. Technical variation and hence the reproducibility of results between different laboratories would be minimized by using consistent computational preprocessing, definitions of features, normalizations, and reporting. This harmonization is especially crucial with respect to multimodal AI models.

- **Explainable and Clinically Interpretable AI**

Future models should deliver meaningful insights into the molecule features that are responsible for classification. Feature-importance methods and SHAP-based analyses are example techniques of explainable AI that can be used to link model predictions to relevant biomarkers in the biological domain. But interpretability needs to be complemented by experimental and clinical evidence to

distinguish between predictive association and causal mechanism.

- **Integration of Multi-Omics and Multimodal Data**

The addition of ctDNA/cfDNA, circulating RNA, proteins, and other circulating markers and clinical variables may enhance the biological representation of early disease. Multi-omics models may be able to integrate complementary signals in the genome, epigenome, transcriptome, and functionality. Practical assay requirements, dimensionality, missing data, feature imbalance, and overfitting, however, need to be controlled through integration strategies.

- **Prospective and Longitudinal Studies**

Future studies are required to assess performance in target populations to be screened. Repeated sampling longitudinally may also confirm if molecular changes in the bloodstream occur before clinical diagnosis and if repeated testing increases the diagnostic sensitivity. These studies should analyze classification performance, as well as false positives, false negatives, downstream diagnostic procedures, and clinical outcomes.

- **Toward Personalized and Population-Level Early Detection**

Use of AI-powered liquid biopsy might aid in population screening as well as individual risk assessment. Future systems will include longitudinal molecular profiles along with demographic and clinical details to identify those that need further investigation. The translation will depend on the evidence of clinical utility, reproducibility, cost-effectiveness, access to the results, and suitable pathways of management after positive or negative findings. Therefore, efforts should focus on developing clinically validated and biologically interpretable systems in the future rather than on algorithmic accuracy.

- **Conclusion**

Combining highly sensitive molecular analyses with computational algorithms that can detect intricate patterns in the circulating markers, AI-driven liquid biopsy is an emerging technology that holds promise for early cancer detection. ctDNA/cfDNA, circulating RNA, miRNAs, and proteins offer complementary information across genomic, epigenetic, transcriptional, regulatory, and functional biological processes. AI techniques can contribute to feature selection, biomarker-signature development, classification, and the integration of these heterogeneous molecular signals, which might enhance the capability to detect weak cancer-associated patterns that can be difficult to recognize with using individual biomarkers.

The evidence reviewed includes evidence of use for individual cancer types as well as multi-cancer early detection. Multimodal and multi-omics approaches are especially relevant since early malignancies could manifest with a range of signals on different molecular layers. Diagnostic performance, however, is reported with some caution, as studies can vary in cancer stage, cohort makeup, biomarker platform, analysis methods and protocol, and validation approaches. A high performance in retrospective or case-control settings does not prove effectiveness in screening populations of asymptomatic individuals.

There are several obstacles to be overcome before it can be widely adopted in clinical practice. Model reliability can be impacted by low abundance of signals from the tumor, biological heterogeneity, pre-analytical and analytical variability, batch effects, high dimensionality, overfitting, limited datasets, population bias, and lack of external validation. To overcome these limitations, standardized workflows and multiple multicenter cohorts are needed, interpretable AI models, rigorous external validation, and prospective longitudinal studies that not only measure the diagnostic performance of the AI models but also the downstream clinical impact.

In conclusion, the future of AI-driven liquid biopsy lies in the intersection of biological understanding and sound computational techniques, along with rigorous clinical validation. Good early-detection systems will probably need reproducible sets of multiple molecular signals, instead of one. To make liquid-biopsy signatures derived from AI tools useful for cancer detection in real-life scenarios, they must be clinically useful, scalable, affordable, and accurate in a representative population.

References

1. J. D. Schiffman, P. G. Fisher, and P. Gibbs, “Early detection of cancer: Past, present, and future,” *ASCO Educational Book*, vol. 35, no. 1, pp. 57–65, 2015.
2. A. A. Ahmed and M. Abedalthagafi, “Cancer diagnostics: The journey from histomorphology to molecular profiling,” *Oncotarget*, vol. 7, no. 36, p. 58696, 2016.
3. R. C. Fitzgerald *et al*., “The future of early cancer detection,” *Nature Medicine*, vol. 28, no. 4, pp. 666–677, 2022.

4. W. She, Y. Garitaonaindia, and Y. Lin, "The latest advances in liquid biopsy for lung cancer—a narrative review," *Translational Lung Cancer Research*, vol. 13, no. 11, p. 3241, 2024.
5. X. Wu *et al*., "Circulating tumor DNA as an emerging liquid biopsy biomarker for early diagnosis and therapeutic monitoring in hepatocellular carcinoma," *International Journal of Biological Sciences*, vol. 16, no. 9, p. 1551, 2020.
6. S. N. Lone *et al*., "Liquid biopsy: A step closer to transform diagnosis, prognosis and future of cancer treatments," *Molecular Cancer*, vol. 21, no. 1, p. 79, 2022.
7. S. Grabuschnig *et al*., "Putative origins of cell-free DNA in humans: A review of active and passive nucleic acid release mechanisms," *International Journal of Molecular Sciences*, vol. 21, no. 21, p. 8062, 2020.
8. F. Leung *et al*., "Circulating tumor DNA as a cancer biomarker: Fact or fiction?," *Clinical Chemistry*, vol. 62, no. 8, pp. 1054–1060, 2016.
9. Y. M. D. Lo *et al*., "Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies," *Science*, vol. 372, no. 6538, Art. no. eaaw3616, 2021.
10. S.-P. Liu *et al*., "The messenger RNA and long non-coding RNA expression profiles in ectopic and eutopic endometrium provide novel insights into endometriosis," *Reproductive and Developmental Medicine*, vol. 3, no. 1, pp. 11–17, 2019.
11. P. S. Mitchell *et al*., "Circulating microRNAs as stable blood-based markers for cancer detection," *Proceedings of the National Academy of Sciences*, vol. 105, no. 30, pp. 10513–10518, 2008.
12. K. Felekkis and C. Papaneophytou, "The circulating biomarkers league: Combining miRNAs with cell-free DNAs and proteins," *International Journal of Molecular Sciences*, vol. 25, no. 6, p. 3403, 2024.
13. V. Ungerer, A. J. Bronkhorst, and S. Holdenrieder, "Preanalytical variables that affect the outcome of cell-free DNA measurements," *Critical Reviews in Clinical Laboratory Sciences*, vol. 57, no. 7, pp. 484–507, 2020.
14. W. Hemstrom *et al*., "Next-generation data filtering in the genomics era," *Nature Reviews Genetics*, vol. 25, no. 11, pp. 750–767, 2024.
15. J. A. B. Brandão, "Dimensionality reduction in survival models based on gene expression data: An application to brain cancer," 2024.
16. Q. Li, Y. He, and J. Pan, "CrossFuse-XGBoost: Accurate prediction of the maximum recommended daily dose through multi-feature fusion, cross-validation screening and extreme gradient boosting," *Briefings in Bioinformatics*, vol. 25, no. 1, Art. no. bbad511, 2024.
17. K. Wang *et al*., "Liquid biopsy techniques and pancreatic cancer: Diagnosis, monitoring, and evaluation," *Molecular Cancer*, vol. 22, no. 1, p. 167, 2023.
18. L. K. K. Rajendran, "Genomic profiling: Utilizing multi-omics data to identify potential therapeutic targets and resistance markers," *Power System Protection and Control*, vol. 52, no. 4, pp. 159–166, 2024.
19. M. Greenacre *et al*., "Principal component analysis," *Nature Reviews Methods Primers*, vol. 2, no. 1, p. 100, 2022.
20. J. Doroszkiewicz, M. Groblewska, and B. Mroczko, "Molecular biomarkers and their implications for the early diagnosis of selected neurodegenerative diseases," *International Journal of Molecular Sciences*, vol. 23, no. 9, p. 4610, 2022.
21. J. E. Medina *et al*., "Cell-free DNA approaches for cancer early detection and interception," *Journal for Immunotherapy of Cancer*, vol. 11, no. 9, Art. no. e006013, 2023.
22. M. Karimzadeh *et al*., "Deep generative AI models analyzing circulating orphan non-coding RNAs enable accurate detection of early-stage non-small cell lung cancer," *medRxiv*, 2024.
23. P. Tighe *et al*., "Utility, reliability and reproducibility of immunoassay multiplex kits," *Methods*, vol. 61, no. 1, pp. 23–29, 2013.
24. Y. Nam *et al*., "Harnessing artificial intelligence in multimodal omics data integration: Paving the path for the next frontier in precision medicine," *Annual Review of Biomedical Data Science*, vol. 7, no. 1, pp. 225–250, 2024.
25. S. Hieke *et al*., "Integrating multiple molecular sources into a clinical risk prediction signature by extracting complementary information," *BMC Bioinformatics*, vol. 17, no. 1, p. 327, 2016.
26. R. Li *et al*., "Machine learning meets omics: Applications and perspectives," *Briefings in Bioinformatics*, vol. 23, no. 1, Art. no. bbab460, 2022.
27. M. Kim and J. Costello, "DNA methylation: An epigenetic mark of cellular memory," *Experimental & Molecular Medicine*, vol. 49, no. 4, Art. no. e322, 2017.

28. L. Pikor *et al*., "The detection and implication of genome instability in cancer," *Cancer and Metastasis Reviews*, vol. 32, no. 3, pp. 341–352, 2013.

.