



Journal Website:
<https://theamericanjournals.com/index.php/tajir>

Copyright: Original content from this work may be used under the terms of the creative commons attributes 4.0 licence.

Research Article

Artificial Intelligence And Machine Learning In Cancer Biomarker Discovery: Emerging Approaches For Early Diagnosis

Submission Date: November 01, 2023, Accepted Date: November 17, 2023,
Published Date: November 30, 2023 |

 Shakhnoza Khamidova
Independent Researcher, USA

ABSTRACT

Early detection is an important component of improving cancer diagnosis and treatment outcomes, yet conventional biomarker approaches often face limitations in sensitivity, specificity, and biological heterogeneity. The increasing availability of high-throughput genomic, transcriptomic, proteomic, metabolomic, and circulating biomarker data has created new opportunities for computational approaches to identify more informative diagnostic patterns. This review examines the application of artificial intelligence (AI) and machine learning (ML) to cancer biomarker discovery, with particular emphasis on laboratory-based approaches for early diagnosis. It discusses major biomarker classes, including circulating proteins, genomic and DNA alterations, microRNAs, metabolites, and liquid-biopsy markers, and examines how ML methods can support feature selection, dimensionality reduction, classification, and integration of molecular measurements. Applications across breast, lung, colorectal, pancreatic, ovarian, prostate, gastric, and liver cancers demonstrate how computational analysis can identify multi-biomarker signatures from complex biological datasets. The review also considers diagnostic performance measures, ROC analysis, validation strategies, overfitting, and the distinction between statistical discrimination and clinical utility. Challenges related to dataset heterogeneity, laboratory variability, external validation, interpretability, reproducibility, data privacy, and clinical implementation remain important barriers to translation. Future progress is likely to depend on integrating multiple molecular and clinical data types through multimodal and multi-omics AI approaches. Ultimately, the clinical value of AI-assisted biomarker discovery will depend on rigorous validation and successful translation of computationally identified signatures into reproducible, practical, and clinically useful laboratory tests.

KEYWORDS

Cancer biomarkers; Artificial intelligence; Machine learning; Early cancer detection; Biomarker discovery; Molecular diagnostics; Liquid biopsy; Multi-omics; Diagnostic signatures; Clinical laboratory medicine; Precision medicine; Deep learning

INTRODUCTION

Burden and challenges of early cancer detection

Cancer remains a major global health burden, and clinical outcomes are closely associated with the stage at which the disease is detected. Identifying malignancy before extensive local invasion or metastatic spread occurs can provide a broader range of treatment options and may improve the chances of achieving long-term disease control. However, detecting cancer at an early stage is challenging because tumors can be biologically heterogeneous and may produce only subtle molecular changes that are difficult to distinguish from normal physiological variation. As a result, many patients may not experience sufficiently specific symptoms during the initial stages of disease, while conventional diagnostic methods may provide clearer evidence only after a tumor has reached a detectable size or caused clinically apparent abnormalities. It is important to distinguish between screening, early detection, and diagnosis, as these terms represent related but distinct stages of cancer management [1]. Screening generally involves testing apparently healthy or asymptomatic populations to identify individuals who may require additional investigation. Early detection refers more broadly to identifying cancer at an earlier point in its natural history, whereas diagnosis involves confirming the presence and characteristics of the disease following clinical suspicion. Biomarkers can support each of these processes, although their requirements vary depending on their intended application. A screening biomarker, for instance, generally requires high sensitivity along with adequate specificity to prevent excessive false-positive results when used in large populations [2]. Diagnostic applications may instead place greater emphasis on distinguishing malignant disease from particular benign or inflammatory conditions.

These differences highlight one of the key challenges in laboratory-based cancer detection. A biomarker may show statistically significant differences between cancer and control groups without providing sufficient diagnostic discrimination for routine clinical use. Variation between patients, differences in tumor subtype and disease stage, age, sex, comorbidities, and sample-processing procedures can all affect measured biomarker concentrations. Therefore, identifying biomarkers that are both biologically meaningful and clinically reliable remains an important challenge in laboratory medicine.

Cancer biomarkers and their role in laboratory medicine

Cancer biomarkers comprise a diverse group of measurable biological characteristics associated with the presence, development, or behavior of malignancy. Conventional laboratory biomarkers include circulating proteins and tumor-associated antigens, while advances in molecular technologies have expanded the field to include genomic alterations, RNA and microRNA expression patterns, DNA methylation, circulating tumor DNA (ctDNA), metabolites, and other molecular features. High-throughput technologies have consequently transformed biomarker research from the investigation of a relatively small number of candidate molecules toward the analysis of thousands of molecular features simultaneously [3]. Protein biomarkers remain important in clinical oncology because they can often be measured using established laboratory platforms. However, their diagnostic performance may be limited when individual proteins are influenced by non-malignant conditions or when their concentrations overlap considerably between cancer and control populations. Molecular biomarkers provide additional information. Genomic alterations can reflect tumor-

specific changes, whereas RNA and microRNA profiles may capture changes in gene regulation. Similarly, ctDNA provides an opportunity to examine tumor-derived genetic material circulating in blood, potentially enabling minimally invasive assessment of malignancy. Metabolomic and epigenetic measurements provide further layers of information concerning altered cellular metabolism and gene regulation [4].

A major limitation of relying on a single biomarker is that cancer is not a uniform biological process. Tumors can differ substantially between patients and even within the same tumor, producing considerable molecular heterogeneity. As a result, a marker that performs well in one population or cancer subtype may have weaker performance in another. This has encouraged increasing interest in **composite biomarker panels**, in which several molecular features are evaluated together. Previous work examining clinically validated omics-based biomarker signatures has emphasized the importance of appropriate study design, feature selection, biological knowledge, and multicohort validation when developing such signatures.

Emergence of AI and machine learning in biomedical research

The growing scale and complexity of laboratory data have made computational approaches increasingly relevant to biomedical research. Artificial intelligence (AI) refers broadly to computational techniques developed to perform tasks that involve aspects of human reasoning, while machine learning (ML) focuses on methods that learn patterns and relationships directly from data [5]. Deep learning (DL), a subfield of ML, uses multilayer neural networks to learn increasingly complex representations from large and diverse datasets. In biomarker research, supervised learning can be applied when samples have predefined outcomes, such as cancer and non-cancer classifications. Algorithms such as logistic regression, support vector machines, random forests, and gradient-boosting methods can identify relationships between molecular features and clinical outcomes. In contrast, unsupervised learning does not depend on predefined labels and can help identify clusters, molecular subtypes, or patterns that may not be apparent within complex

datasets [6]. DL methods can further support the analysis of high-dimensional molecular data and datasets containing different types of biological information. These approaches are particularly relevant to multi-omics research. Genomic, transcriptomic, proteomic, and metabolomic datasets may contain thousands of potential features, making conventional one-variable-at-a-time analysis inadequate for capturing interactions and complex relationships among them. Recent reviews of ML-based multi-omics analysis have highlighted the increasing application of both supervised and unsupervised methods for cancer classification, biomarker discovery, and molecular characterization.

Rationale for AI-driven biomarker discovery

The central rationale for applying AI and ML to cancer biomarker discovery is therefore not simply to automate laboratory diagnosis, but to identify clinically meaningful patterns within increasingly complex biological datasets. ML methods can assist in reducing high-dimensional datasets to informative features, recognizing nonlinear relationships, prioritizing candidate biomarkers, classifying samples according to disease status, and constructing multivariate diagnostic signatures. Deep-learning approaches have also demonstrated potential for integrating different molecular data types and identifying predictive biomarker patterns that may be difficult to detect using conventional analytical approaches alone [7].

Nevertheless, computational discovery should not be considered equivalent to clinical validation. A biomarker signature identified from a retrospective dataset must undergo independent validation, analytical assessment, and ultimately clinical evaluation before it can be incorporated into routine laboratory practice. Issues including overfitting, data heterogeneity, reproducibility, interpretability, and external validation remain important barriers to translation. Against this background, this review examines the emerging role of AI and ML in cancer biomarker discovery, with particular emphasis on their application to molecular and laboratory-derived data for early cancer detection. It considers major biomarker classes, computational strategies for identifying informative features and

biomarker combinations, cancer-specific applications, and the challenges that must be addressed before AI-derived biomarker signatures can achieve broader clinical implementation.

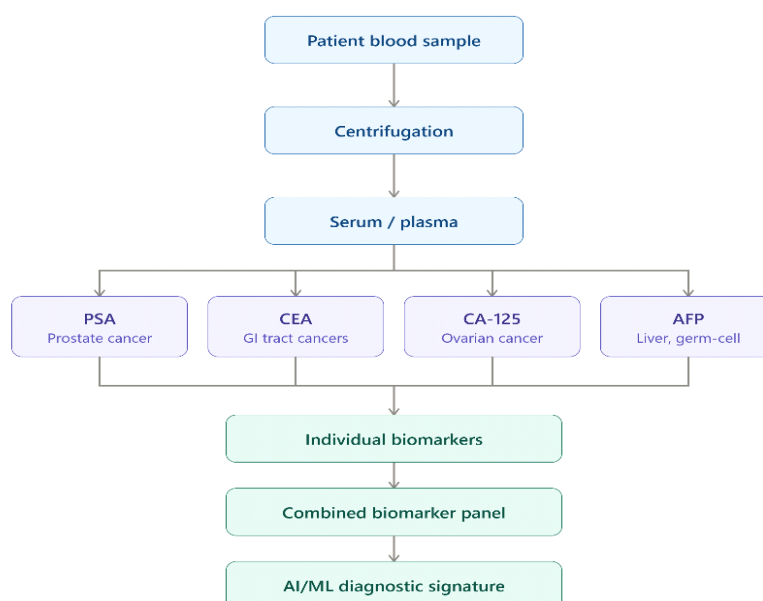
Cancer Biomarkers Relevant to AI-Based Discovery

Cancer biomarker research has expanded considerably with the development of high-throughput molecular and laboratory technologies. Rather than relying solely on a single measurable analyte, contemporary biomarker studies increasingly examine molecular patterns across proteins, nucleic acids and metabolites. This shift is particularly relevant to artificial intelligence (AI)-based discovery because these datasets contain numerous interacting variables that may collectively distinguish malignant from non-malignant states.

Protein and serum biomarkers

Protein-based biomarkers remain an important component of laboratory oncology. Circulating proteins, enzymes and tumor-associated antigens can be measured in serum or plasma using established

biochemical and immunoassay-based methods. Examples include prostate-specific antigen (PSA), carcinoembryonic antigen (CEA), cancer antigen 125 (CA-125) and alpha-fetoprotein (AFP). Their established laboratory availability makes them attractive candidates for computational analysis; however, their diagnostic utility is frequently limited when concentrations overlap between patients with cancer and those with benign or inflammatory conditions [3]. Consequently, an individual serum marker may demonstrate biological association with malignancy without achieving sufficient sensitivity or specificity for population-level early detection. AI-based analysis provides an opportunity to examine several protein measurements simultaneously. Instead of assigning diagnostic significance to one analyte, machine-learning models can identify combinations of protein concentrations that collectively distinguish disease states. This approach is particularly relevant when relatively modest changes across several analytes may contain more diagnostic information than a marked alteration in a single biomarker [4].



Circulating protein biomarkers and their integration into AI-assisted cancer biomarker discovery.

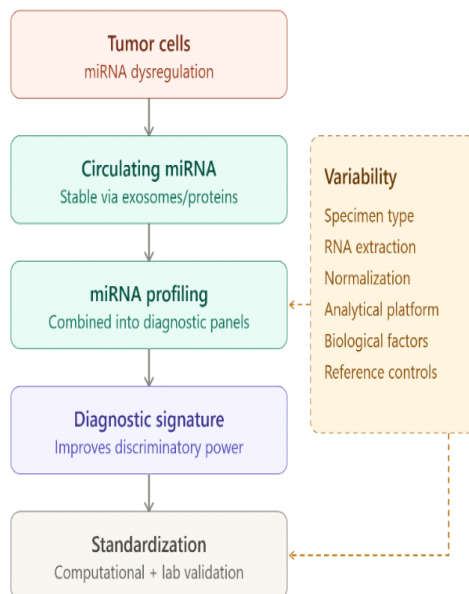
Blood-derived serum or plasma contains multiple measurable proteins and tumor-associated antigens that can provide diagnostic information. Established examples include PSA, CEA, CA-125, and AFP. Rather than relying exclusively on individual biomarkers, machine-learning approaches can integrate multiple circulating measurements to identify composite diagnostic signatures.

Genomic and DNA-based biomarkers

Genomic biomarkers provide a more direct representation of molecular alterations associated with malignant transformation. Somatic mutations, copy-number alterations and aberrant DNA methylation are among the molecular features investigated for cancer detection and characterization. High-throughput sequencing has enabled simultaneous examination of numerous genomic regions, creating datasets that are well suited to computational feature selection and classification [8]. Circulating tumor DNA (ctDNA), a fraction of cell-free DNA released into the circulation by tumor cells, has attracted particular interest as a minimally invasive biomarker. ctDNA may contain tumor-specific mutations, copy-number changes, methylation abnormalities and other molecular characteristics that can potentially distinguish malignant from non-malignant biological material. However, its concentration can be extremely low in early disease, making analytical sensitivity a major consideration. Earlier evaluations of ctDNA for early cancer detection identified substantial pre-analytical and analytical challenges and emphasized the need for larger validation studies before routine screening applications could be established.

RNA and microRNA biomarkers

RNA-based biomarkers provide another molecular layer for cancer detection. Among these, microRNAs (miRNAs) have received considerable attention because these small non-coding RNAs regulate gene expression and may become dysregulated during malignant transformation. miRNAs can also be detected in circulation, where they may remain relatively stable through association with proteins or extracellular vesicles. This stability has contributed to interest in circulating miRNAs as minimally invasive biomarkers. Circulating miRNA profiles may differ according to tumor type, disease stage and biological characteristics, creating the possibility of using expression signatures rather than individual miRNAs for diagnosis. For example, a 2022 systematic review and meta-analysis identified several circulating miRNAs with potential diagnostic value in breast cancer and demonstrated that combining selected miRNAs into panels could improve discriminatory performance compared with individual markers [9]. Despite this potential, clinical translation remains challenging. Differences in specimen type, sample handling, RNA extraction, normalization strategies and analytical platforms can influence measured miRNA concentrations. Biological variation and the lack of universally standardized reference controls can also contribute to inconsistent findings between studies. These limitations make circulating miRNAs an important example of a biomarker class in which computational methods may assist in identifying reproducible combinations of molecular features while laboratory standardization and independent validation remain essential.



miRNA Biomarker Pipeline

Metabolic biomarkers

Metabolomics offers another important source of biomarkers by characterizing small molecules that reflect alterations in cellular metabolism. Cancer-associated metabolic reprogramming can produce measurable changes in amino acids, lipids, organic acids and other metabolites in blood, serum, plasma or urine. Consequently, metabolic profiles may provide information that is complementary to protein and nucleic-acid biomarkers. Studies have demonstrated the potential of serum and urine metabolomic signatures for distinguishing malignant from non-malignant conditions, including cancers that are difficult to identify at an early stage [10]. However, metabolomic investigations typically generate large numbers of correlated features, making feature selection and interpretation challenging. Machine-learning approaches can assist in identifying combinations of metabolites that provide greater discriminatory information than individual metabolic markers.

Composite and multi-biomarker signatures

The limitations of individual biomarkers have progressively shifted cancer biomarker research from

single biomarkers toward biomarker panels and computational signatures. A single marker may reflect only one aspect of tumor biology, whereas combining several molecular features can capture complementary characteristics of malignant disease [11]. Such combinations may include proteins, genomic alterations, miRNAs or metabolites, depending on the cancer and available laboratory data. Computational biomarker signatures extend this concept by using statistical or machine-learning methods to select informative features from a larger molecular dataset and combine them into a diagnostic model. This approach is particularly relevant to early cancer detection, where disease-associated molecular changes may be subtle and distributed across multiple biological pathways. Nevertheless, increasing the number of candidate biomarkers does not inherently improve diagnostic performance. Appropriate feature selection, avoidance of overfitting, independent validation and assessment in clinically representative populations remain necessary before a computational signature can be considered clinically useful.

Machine Learning Approaches for Cancer Biomarker Discovery

The increasing use of high-throughput sequencing, proteomics, metabolomics and other molecular profiling technologies has generated datasets containing substantially more variables than can be evaluated effectively through conventional analytical approaches alone. In cancer biomarker research, this high dimensionality creates both an opportunity and a methodological challenge: potentially informative molecular patterns may be present within the data, but irrelevant, redundant or correlated features can obscure these patterns. Machine learning (ML) provides a collection of computational approaches for preprocessing, reducing, classifying and integrating such data, thereby supporting the identification of candidate biomarkers and multivariate diagnostic signatures.

Data preprocessing and normalization

The reliability of an ML-derived biomarker is strongly dependent on the quality of the data supplied to the model. Molecular datasets commonly contain missing observations arising from technical limitations, insufficient specimen quantity or assay-related factors [12]. Depending on the extent and mechanism of missingness, researchers may use appropriate imputation strategies or exclude variables with excessive missing values. Outlier detection is similarly important because extreme measurements may represent genuine biological variation, technical error or pre-analytical problems and should not be removed indiscriminately.

Normalization and scaling are particularly important when biomarkers are measured on different numerical scales. Transformation and normalization procedures can reduce unwanted variation and make measurements more comparable across samples. In high-throughput studies, batch effects represent an additional concern because samples processed at different times, laboratories or experimental batches may exhibit systematic differences unrelated to the underlying disease state. If these effects are not appropriately controlled, an ML model may learn technical differences rather than biologically meaningful characteristics [13]. Accordingly, quality control and

preprocessing should precede model development and should be designed to preserve genuine biological variation while minimizing technical artifacts.

Feature selection and dimensionality reduction

Feature selection is particularly important in omics-based biomarker research because the number of measured genes, proteins, metabolites or other molecular variables can greatly exceed the number of patient samples. Retaining all available features can increase computational complexity and encourage overfitting, whereas selecting informative variables can produce more interpretable and potentially more robust models. Feature-selection methods are commonly divided into filter, wrapper and embedded approaches. Filter methods evaluate features according to statistical or information-based characteristics independently of a final predictive model. Wrapper methods assess subsets of variables according to their contribution to a selected learning algorithm, while embedded methods perform feature selection during model training. LASSO regression is an example of an embedded approach that applies regularization to shrink less informative coefficients toward zero, thereby producing a more compact set of candidate predictors [14].

Recursive feature elimination (RFE) provides another strategy by repeatedly removing the least informative variables according to model-derived importance. In contrast, principal component analysis (PCA) transforms correlated variables into a smaller number of orthogonal components that capture major sources of variation. PCA is primarily a dimensionality-reduction technique rather than a direct biomarker-selection method, but it can assist visualization, noise reduction and subsequent modelling. Reviews of successful omics-based biomarker studies emphasize the importance of early filtering, dimensionality reduction, rigorous model development and independent validation.

Supervised learning

Supervised learning is particularly suitable when molecular profiles are associated with known clinical outcomes, such as cancer versus non-cancer status or different disease subtypes. Logistic regression provides a relatively interpretable baseline for estimating the probability of a diagnostic outcome from selected

biomarkers. Its simplicity can be advantageous when the objective is to construct a clinically understandable biomarker model. Support vector machines (SVMs) can identify complex decision boundaries and have been widely applied to high-dimensional molecular classification problems. Decision trees provide rule-based classification and can capture nonlinear relationships, although individual trees may be sensitive to changes in the training data. Random forests address some of these limitations by combining multiple decision trees and can additionally provide measures of feature importance, making them useful for prioritizing candidate biomarkers.

Gradient-boosting approaches construct a sequence of predictive models in which subsequent models focus on errors made by earlier ones. XGBoost, an optimized gradient-boosting implementation, has consequently become a useful method for structured biomedical datasets. These algorithms do not merely classify samples; when appropriately interpreted, their feature-selection or importance measures can help identify molecular variables associated with diagnostic discrimination. Nevertheless, predictive importance should not automatically be interpreted as biological causation, and candidate biomarkers identified computationally require independent laboratory and clinical validation.

Unsupervised learning

Unlike supervised approaches, unsupervised learning does not require predefined disease labels. This makes it valuable when the molecular structure of a dataset is not completely understood. K-means clustering can group samples according to similarities across selected molecular measurements, whereas hierarchical clustering produces a nested representation of relationships among samples or features. Such methods can reveal molecular subgroups that may not correspond to conventional clinical classifications. PCA

can also be used to visualize major patterns of variation and determine whether samples naturally separate according to disease status or other characteristics. In cancer research, unsupervised methods are particularly relevant to molecular subtyping because tumors classified under the same clinical diagnosis may exhibit substantially different molecular profiles. Multi-omics studies have demonstrated the usefulness of unsupervised learning for identifying disease subtypes, molecular modules and potential biomarker patterns.

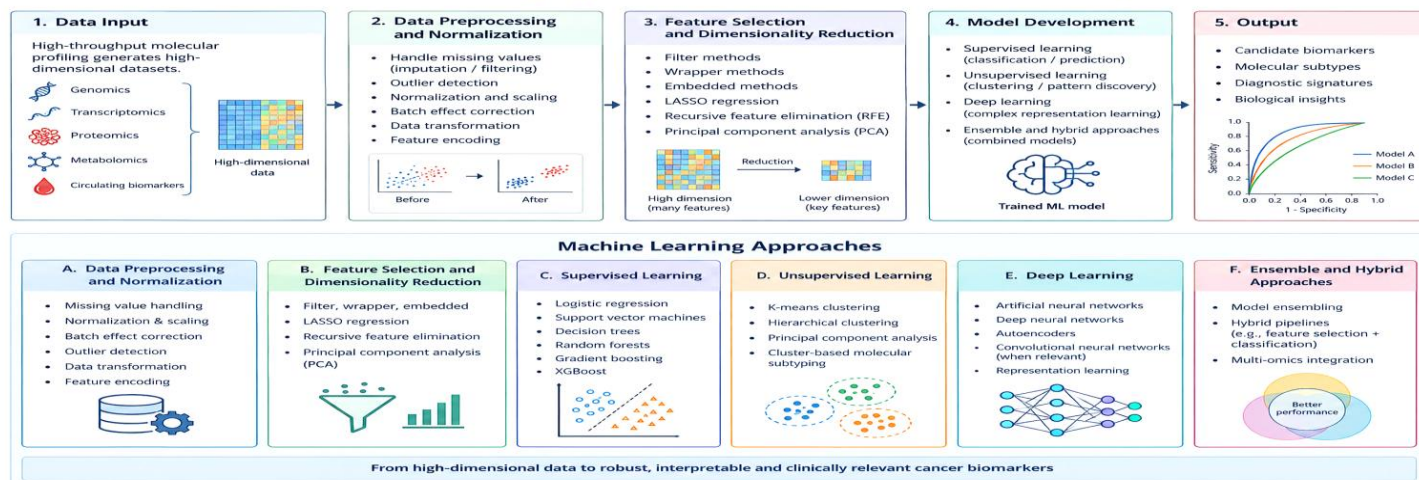
Deep learning

Deep learning (DL) extends ML through multilayer neural-network architectures capable of learning complex representations from high-dimensional data. Artificial neural networks and deeper neural architectures can model nonlinear relationships among molecular features, while autoencoders can learn compact representations of high-dimensional omics data and are therefore useful for dimensionality reduction and feature extraction. Convolutional neural networks (CNNs) have a more established role in image-based oncology, but they have also been incorporated into multimodal approaches in which pathology or other imaging information is combined with molecular data. Such approaches illustrate how molecular biomarkers can potentially be interpreted alongside other biological information [15].

Ensemble and hybrid approaches

No single algorithm is universally optimal for cancer biomarker discovery because molecular datasets differ in dimensionality, sample size, biological structure and measurement characteristics. Ensemble approaches therefore combine multiple models to improve stability, classification performance or feature prioritization. Random forests represent one example, while other ensemble strategies combine predictions from different algorithms.

Machine Learning Approaches for Cancer Biomarker Discovery



Machine Learning Approaches for Cancer Biomarker Discovery

Hybrid approaches can also combine complementary techniques, such as applying feature selection before supervised classification, or using dimensionality reduction together with clustering and subsequent predictive modelling. These strategies have become increasingly common in multi-omics research because different molecular layers provide distinct information about tumor biology. The goal is not simply to achieve the highest possible predictive accuracy, but to develop biomarker signatures that are reproducible, biologically meaningful, and clinically relevant. Overall, ML offers a methodological framework that can be applied across several stages of biomarker discovery, from data quality control and feature reduction to molecular classification and biomarker prioritization. However, its effectiveness in cancer biomarker discovery depends not only on the sophistication of the selected algorithm, but also on appropriate study design, meaningful biological interpretation, rigorous validation, and the quality of the underlying laboratory measurements.

AI-Driven Discovery Across Major Biomarker Types

Proteomic profiling provides a particularly rich source of potential cancer biomarkers because malignant transformation can affect the abundance, modification, and interactions of large numbers of proteins. Serum and plasma are especially useful for biomarker discovery because they can be collected through minimally

invasive sampling and contain proteins released by tumors, surrounding tissues, and systemic responses to disease. Mass spectrometry has greatly increased the number of proteins that can be measured simultaneously, but these datasets also contain a large number of variables whose diagnostic relevance may not be determined reliably through individual statistical comparisons alone. Machine learning can analyze these multidimensional profiles and identify combinations of proteins that distinguish malignant from non-malignant conditions. Deep-learning approaches have also been incorporated into proteomic workflows, where they can predict peptide measurements from amino acid sequences and support the identification and interpretation of mass-spectrometry data. At the biomarker level, supervised learning can then analyze protein abundance profiles to identify features associated with disease status and develop diagnostic signatures. Rather than evaluating each protein independently, these models can capture complementary information from multiple proteins, which is particularly valuable when individual biomarkers have limited specificity. Computational analysis has therefore shifted proteomic biomarker discovery from searching for a single highly expressed tumor-associated protein toward identifying reproducible patterns across multiple proteins.

The clinical relevance of this approach ultimately depends on whether the selected proteins remain informative across independent patient populations. A protein may appear highly discriminative within a discovery cohort because of differences in age, inflammation, sample handling, or disease composition rather than because it reflects a tumor-specific biological signal. Consequently, machine-learning-based proteomic discovery is most useful when computationally selected proteins are subsequently evaluated using independent cohorts and laboratory assays.

AI for Genomic and DNA-Based Biomarker Discovery

Genomic data provide another important source of biomarkers because cancer cells accumulate characteristic mutations, copy-number changes, methylation abnormalities, and other alterations during tumor development. Modern sequencing can identify thousands or millions of genomic variants, creating a feature-selection problem in which only a small proportion may have meaningful diagnostic or biological relevance. Machine learning has been used to distinguish potentially functional cancer-associated alterations from large numbers of passenger mutations [16].

One important application is the identification of cancer driver genes and driver mutations. These alterations contribute directly to tumor development and therefore provide biologically meaningful candidates for disease characterization. Computational models can combine mutation frequency with information about gene function, interaction networks, sequence characteristics, and other genomic features to estimate which genes or mutations are likely to contribute to tumorigenesis [17]. Machine-learning and deep-learning approaches have consequently been applied to prioritize cancer driver genes from large collections of somatic mutations, helping reduce the search space for experimentally relevant biomarkers. DNA methylation provides another promising feature because abnormal methylation patterns can occur early during malignant transformation. Instead of examining a single methylated locus, machine-learning models can analyze methylation across numerous genomic regions and

identify combinations associated with particular tumor types. Such approaches are particularly relevant to blood-based detection, where tumor-derived DNA may constitute only a small fraction of circulating cell-free DNA. The ability to combine multiple weak genomic signals is therefore an important advantage of computational analysis.

AI for microRNA Biomarker Discovery

MicroRNAs represent another class of circulating biomarkers that can be investigated using machine learning. These short non-coding RNAs regulate gene expression and can exhibit altered expression patterns during tumor development. Because circulating microRNAs can be detected in blood and other biological fluids, their expression profiles have been investigated as potential minimally invasive indicators of cancer.

The value of machine learning becomes evident when multiple microRNAs are considered simultaneously. Instead of selecting biomarkers solely according to whether an individual microRNA is significantly different between patients and controls, computational models can evaluate combinations of expression levels and determine which features contribute most strongly to classification. Longitudinal measurements can provide an additional dimension. For example, a validation study involving serum samples from patients with colon, lung, and breast cancer examined 21 circulating microRNAs using computational models. The analysis distinguished static biomarkers, based on differences between cancer and control samples, from dynamic biomarkers, in which changes in expression over time provided additional diagnostic information. miR-99a-5p showed relevance across the three cancer groups, while miR-155-5p was associated with the prediction of lung and colon cancers [18]. A gradient-boosted-tree model achieved an average classification accuracy of 79.9% for cancer versus non-cancer samples. These findings illustrate why longitudinal and multivariable analysis can be useful in circulating RNA research. The diagnostic information may not reside exclusively in the absolute concentration of a single microRNA; it may also emerge from relationships among several microRNAs or from changes observed over time.

AI for Metabolomic Biomarker Discovery

Cancer-associated metabolic reprogramming produces measurable changes in small molecules such as amino acids, lipids, organic acids, and other metabolites. Metabolomics can therefore provide a biochemical representation of physiological changes associated with malignant disease. Techniques such as mass spectrometry and nuclear magnetic resonance generate profiles containing hundreds or thousands of metabolic features, making computational pattern recognition particularly valuable. Machine learning has been used to transform these profiles into diagnostic or biomarker signatures. In a study of non-small-cell lung cancer, untargeted serum metabolomics identified 278 altered metabolic features between patients and healthy controls. Machine-learning analysis of these features produced a panel of 10 metabolic biomarkers with an area under the ROC curve of 0.95 in the validation cohort, including an AUC of 0.95 for early-stage disease detection [19]. The altered metabolites were associated with pathways including phenylalanine metabolism, linoleic acid metabolism, and bile-acid biosynthesis, linking the computational signature with biologically interpretable metabolic changes. Metabolomic analysis has also been combined with other clinical and laboratory measurements. For example, NMR-based serum metabolomics combined with hormone-receptor status, HER2, and Ki-67 was used with machine learning to predict response to neoadjuvant chemotherapy in breast cancer. Leucine, formate, valine, and proline were among the metabolites contributing to discrimination between treatment-response groups, while the integrated model achieved more than 80% accuracy [20]. The diagnostic potential of metabolomics is therefore not limited to identifying individual metabolites. AI can identify coordinated metabolic patterns that reflect altered biochemical pathways and can combine these patterns with established laboratory or clinical variables.

AI for Liquid-Biopsy Biomarkers

Liquid biopsy is one of the most direct applications of AI-assisted biomarker discovery because blood and other body fluids contain a range of molecular traces released by tumors, including circulating tumor DNA (ctDNA), cell-free DNA (cfDNA), circulating RNA, extracellular

vesicles, circulating tumor cells, and tumor-associated proteins [21]. Each of these components reflects different aspects of tumor biology, although their concentrations can be extremely low, particularly when the tumor burden is limited. Machine learning makes it possible to analyze multiple subtle molecular signals together. For instance, cfDNA can be examined not only for tumor-specific sequence alterations but also for patterns of DNA methylation, fragment-size distributions, nucleosome positioning, and other biological characteristics. These features can provide information about both the presence of malignancy and the tissue from which the abnormal DNA originated. Computational models are particularly useful in this context because the diagnostic signal may be spread across numerous weakly informative features rather than concentrated in a single mutation. Liquid-biopsy research has also combined different circulating components within machine-learning workflows. Studies using blood, urine, and saliva have examined circulating tumor cells, cfDNA, ctDNA, cell-free RNA, exosomes, and tumor-educated platelets as potential inputs for cancer detection [22]. These approaches have been investigated across several cancer types, including pancreatic, hepatic, breast, and oral cancers. Combining multiple liquid-biopsy components can therefore provide a more comprehensive molecular representation of disease than relying on a single circulating biomarker.

Particularly important is the development of cfDNA analysis beyond conventional mutation detection. Methylation patterns, fragmentomics, and nucleosome-related features can provide tissue-specific fingerprints, while machine-learning models can integrate these features to distinguish cancer-associated cfDNA from background circulating DNA. This creates a pathway toward blood-based detection based on molecular patterns rather than requiring a single high-abundance tumor mutation.

AI for Multi-Biomarker and Integrated Molecular Signatures

The most significant development in AI-based biomarker discovery is the movement from individual biomarkers toward integrated molecular signatures. Cancer is

biologically heterogeneous, and no single biomarker is consistently expressed across all patients or disease stages. A protein may capture an inflammatory response, a mutation may reflect tumor genetics, a microRNA may indicate altered gene regulation, and a metabolite may represent changes in cellular metabolism. Combining these different forms of information can therefore provide a more comprehensive representation of malignancy. Machine learning provides several ways to integrate such heterogeneous measurements. Features from proteomics, genomics, transcriptomics, metabolomics, and clinical laboratory data can be combined into a unified predictive model, or separate models can first be developed for individual data types and subsequently integrated. Multi-omics approaches have been used to identify molecular signatures associated with cancer subtypes, patient stratification, diagnosis, and prognosis [23]. Computational resources developed for cancer biomarker discovery have similarly incorporated multi-omics repositories, analytical tools, and machine-learning methods for diagnostic and predictive biomarker development. An important advantage of integrated signatures is that complementary biological information can compensate for weaknesses in individual biomarkers. For example, combining circulating protein measurements with genomic or metabolic features may improve discrimination when none of the individual measurements provides sufficient specificity. AI-based feature selection can also reduce a large multi-omics dataset to a smaller set of biologically informative variables that can potentially be measured through practical laboratory assays.

However, increasing the number of biomarkers does not automatically produce a better diagnostic test. A highly complex signature can fit characteristics specific to the discovery cohort rather than general biological patterns. For this reason, the most useful AI-derived signatures are those that maintain diagnostic performance across independent cohorts and can ultimately be translated into reproducible laboratory measurements. The goal of computational biomarker discovery is therefore not simply to generate large molecular signatures, but to identify a sufficiently informative and biologically

meaningful combination of markers that can support reliable cancer detection.

Cancer-Specific Applications of AI-Based Biomarker Discovery

The usefulness of AI-based biomarker discovery becomes clearer when examined across individual cancer types, where differences in tumor biology, circulating biomarker abundance, and clinical presentation create distinct diagnostic challenges. Applications in breast, lung, colorectal, pancreatic, ovarian, prostate, gastric, and liver cancers illustrate how machine learning can extract clinically relevant patterns from proteins, nucleic acids, metabolites, and combinations of laboratory measurements.

Breast Cancer

Breast cancer has been extensively investigated using circulating molecular biomarkers, particularly proteins and microRNAs. Blood-based microRNA profiling has shown that early breast cancer can produce detectable changes in circulating RNA patterns [24]. For example, profiling of whole blood identified dozens of differentially expressed microRNAs between women with early-stage breast cancer and healthy controls, with both increased and decreased expression contributing to the molecular profile. Subsequent validation of selected microRNAs demonstrated the potential for reducing a large expression dataset to a smaller diagnostic signature. Machine-learning analysis further supports the use of combinations rather than isolated microRNAs. A plasma-based two-microRNA signature consisting of miR-30b-5p and miR-99a-5p was constructed using dimensionality-reduction analysis and subsequently evaluated in an independent cohort. The two markers provided complementary information and demonstrated potential for detecting very early breast cancer, illustrating how computational feature reduction can convert a large circulating RNA profile into a compact diagnostic signature.

Lung Cancer

Lung cancer provides an important example of how AI can support the analysis of blood-based metabolites. Since early-stage disease may cause only subtle biochemical changes throughout the body, metabolomic profiling can detect alterations that may

be difficult to identify through conventional single-marker testing. In one plasma metabolomics study, 61 metabolites were measured using targeted LC-MS/MS, and machine-learning analysis identified a combination of six metabolites that could distinguish stage I lung cancer from healthy individuals. The resulting signature achieved an AUC of 0.989 [25], indicating that the selected metabolic features provided considerably more discriminatory information when considered together than when individual metabolites were assessed separately. Protein-based liquid-biopsy approaches have also examined molecular differences between lung cancer and benign or healthy conditions. Exosomal proteomic analysis has been used to identify protein panels associated with lung adenocarcinoma and lung squamous-cell carcinoma. Machine-learning-based feature selection generated separate multi-protein signatures for these histological subtypes, with the signatures maintaining discriminatory ability in early-stage disease. This example demonstrates how AI can combine several moderately informative proteins to produce a more specific molecular classification than conventional approaches based on individual tumor markers.

Colorectal Cancer

Colorectal cancer is well suited to metabolomic biomarker research because malignant transformation is associated with changes in glucose, amino-acid, lipid, and other metabolic pathways. Machine learning can capture these biochemical changes collectively and use them as a metabolic fingerprint of the disease. Plasma metabolic profiling using mass spectrometry has identified colorectal-cancer-specific metabolic patterns containing multiple potential biomarkers. In one study, a 34-metabolite signature was derived from metabolic fingerprints, and the resulting machine-learning classifier achieved AUC values of 0.947 in the discovery cohort and 0.912 during independent validation. DNA methylation offers another complementary source of molecular information. Abnormal methylation of genes involved in transcriptional regulation and tumor development can occur in colorectal cancer and may be detected in DNA obtained from blood samples. Computational analysis of methylation profiles has

therefore been used to reduce large groups of candidate genes to smaller combinations that can be applied for diagnostic classification. Together, metabolic and epigenetic information demonstrate how colorectal cancer detection can move beyond individual biomarkers and toward integrated signatures that combine signals from different molecular classes.

Pancreatic Cancer

Pancreatic ductal adenocarcinoma presents a particularly challenging problem for biomarker discovery because the disease is often detected only after it has reached an advanced stage. At the same time, individual circulating biomarkers generally do not provide sufficient sensitivity or specificity for reliable early detection. These limitations have encouraged researchers to develop predictive models that combine multiple blood-based biomarkers, each of which may provide limited information on its own.

A large serum-based analysis evaluated 31 biologically selected blood analytes, including proteins and other circulating measurements, across healthy individuals, patients with benign pancreatic disease, and patients with early-stage pancreatic cancer. Instead of examining each biomarker independently, the machine learning approach assessed the relationships and combined patterns among changes in these analytes. The resulting classification model achieved an AUC of 0.920 for distinguishing pancreatic cancer from both benign and healthy controls, and 0.944 when compared with healthy controls alone. Similar performance was also observed during independent validation, supporting the consistency of the model across datasets [26]. The finding is important because it demonstrates how individually modest serum biomarkers can collectively provide a substantially stronger diagnostic signal. Serum microRNAs have provided another route for computational pancreatic cancer detection. An artificial neural-network model based on selected serum microRNAs identified three candidate markers miR-4648, miR-125b-1-3p, and miR-3201 and achieved AUC values above 0.92 in training and verification groups. The analysis also separated patients into molecular subgroups, showing that the same computational

framework can provide both diagnostic classification and biological stratification.

Evaluating AI-Derived Cancer Biomarkers

The identification of a promising biomarker through machine learning is only the first stage of diagnostic development. A computational model may identify a molecular pattern that separates cancer from non-cancer samples, but its usefulness depends on how accurately and consistently that pattern performs in clinically relevant populations. Evaluation therefore requires assessment of diagnostic accuracy, discrimination, validation, reproducibility, and practical clinical utility.

Diagnostic Performance

Sensitivity and specificity are fundamental measures for evaluating an AI-derived cancer biomarker. Sensitivity as given in eq. (1) represents the proportion of individuals with the target cancer who are correctly identified by the test, whereas specificity represents the proportion of individuals without the disease who are correctly classified as negative. For early cancer detection, high sensitivity is particularly important because a false-negative result may delay further investigation in a patient with an otherwise potentially treatable disease.

Specificity as given in eq. (2) is equally important in screening because excessive false-positive results can lead to unnecessary imaging, invasive procedures, anxiety, and healthcare expenditure.

where:

- TP = true positives
- TN = true negatives
- FP = false positives
- FN = false negatives

Accuracy, calculated from the proportion of correctly classified individuals among the entire study population, provides a general measure of classification performance but can be misleading when cancer and non-cancer groups are imbalanced. A model tested in a population containing many more healthy individuals than patients with cancer can achieve high accuracy while performing poorly for the minority cancer group.

Positive predictive value (PPV) and negative predictive value (NPV) address a different clinical question: what does a test result mean for an individual patient? PPV represents the probability that a person with a positive result actually has the disease, whereas NPV represents the probability that a person with a negative result is disease-free. Unlike sensitivity and specificity, predictive values depend strongly on disease prevalence. A biomarker that performs well in a high-risk clinical cohort may therefore have a substantially different PPV when applied to a general screening population.

ROC Curves and AUC

Many AI-based diagnostic models produce a continuous probability or score rather than an immediate positive/negative result. A threshold must then be selected to convert the predicted probability into a diagnostic classification. Changing this threshold alters the balance between sensitivity and specificity. A lower threshold may identify more cancers but also increase false-positive results, whereas a higher threshold may improve specificity at the cost of missed cases. A receiver operating characteristic (ROC) curve displays this relationship across possible thresholds by plotting sensitivity against the false-positive rate. The area under the ROC curve (AUC) summarizes the model's ability to discriminate between individuals with and without the target condition across these thresholds. An AUC close to 1 indicates strong discrimination, whereas an AUC near 0.5 indicates performance close to random classification. AUC is useful for comparing biomarker models because it evaluates discrimination across the full range of thresholds rather than relying on one selected cutoff. However, a high AUC does not establish that a biomarker is clinically useful. Two models can have similar AUCs but produce different sensitivities and specificities at the threshold relevant to cancer screening. Threshold selection should therefore reflect the intended clinical application and the relative consequences of false-negative and false-positive results. Diagnostic performance measures are also influenced by patient spectrum and study design, making direct comparison between studies inappropriate when their populations or disease definitions differ substantially.

Key Measures and Considerations for Evaluating AI-Derived Cancer Biomarkers

Evaluation aspect	What it indicates
Sensitivity	Ability to correctly identify patients with cancer
Specificity	Ability to correctly identify individuals without cancer
PPV	Probability that a positive result represents true cancer
NPV	Probability that a negative result represents absence of cancer
AUC	Overall ability of a model to discriminate between cancer and non-cancer
Cross-validation	Estimates model performance while reducing dependence on a single data split
External validation	Tests whether performance generalizes to an independent patient cohort

Internal versus External Validation

Validation helps determine whether the apparent performance of an AI-derived biomarker is likely to remain consistent beyond the data used to develop the model. During model development, the training dataset is used to estimate model parameters and identify the biomarker features that contribute to classification. A separate validation dataset can then be used to tune hyperparameters, compare different models, or determine an appropriate decision threshold. A test dataset that remains completely independent of model development provides a more rigorous assessment of performance. Ideally, the test data should not have contributed to feature selection, preprocessing decisions, model tuning, or threshold determination. If it does, information from the test set can indirectly influence the development process and result in an overly optimistic estimate of model performance.

External validation provides another, particularly important, level of evidence. An AI-derived biomarker developed in one hospital, laboratory, geographic region, or patient population should ideally be evaluated using an independent cohort collected under different conditions. Differences in demographic characteristics, disease prevalence, laboratory platforms, sample handling procedures, recruitment criteria, and clinical practices can affect the distribution of biomarker measurements. A model that performs well only on the dataset used for its development provides limited evidence that it will generalize to other populations or

settings. External validation is therefore an essential step before a diagnostic prediction model can be considered for broader clinical application.

Cross-Validation and Overfitting

Cancer biomarker datasets often contain thousands of molecular features, while the number of available patients is comparatively small. This imbalance creates a considerable risk of overfitting, where a model learns patterns specific to the development dataset rather than patterns that reflect broader characteristics of cancer biology. Cross-validation provides a useful approach for estimating model performance while making efficient use of limited data. In k-fold cross-validation, the dataset is divided into several subsets, with the model trained on part of the data and evaluated on the remaining subset. This process is repeated so that each subset is used as validation data at some point. Repeated cross-validation can provide a more reliable estimate of model performance than relying on a single small hold-out sample. However, cross-validation cannot replace genuine external validation. If feature selection, normalization, or other preprocessing steps are carried out using the complete dataset before cross-validation, information may leak between the training and validation folds, resulting in artificially inflated performance. Similarly, repeatedly testing and tuning a model against the same validation data can gradually cause the model to adapt specifically to that dataset.

Clinical Utility versus Statistical Performance

Statistical discrimination is important, but by itself it is not enough to support clinical implementation. An AI-derived biomarker must be measured reliably under routine laboratory conditions, produce consistent results, and provide results within a clinically acceptable timeframe. The method used to measure the biomarker also affects its practical feasibility. For example, a complex multi-omics signature that depends on expensive mass spectrometry or extensive sequencing may be less accessible than a smaller biomarker panel that can be measured using established laboratory platforms. Clinical utility is also influenced by the population in which the test is applied. A biomarker developed using patients with confirmed cancer and healthy controls may perform differently when applied to individuals with nonspecific symptoms or to asymptomatic populations undergoing screening. Differences in age, sex, ethnicity, comorbidities, medication use, disease stage, and disease prevalence can affect both biomarker distributions and predictive values.

Reproducibility is therefore a fundamental requirement for clinical translation. The laboratory assay, preprocessing pipeline, selected features, model parameters, and decision threshold should be described clearly enough to allow independent evaluation. Transparent reporting of prediction-model development and validation is also important for evaluating methodological quality and potential clinical usefulness. The TRIPOD framework emphasizes the complete reporting of model development, validation, predictors, participants, and performance measures, helping researchers assess these aspects more consistently. Ultimately, an AI-derived cancer biomarker should move through a defined process from computational discovery to analytical verification, independent validation, and evaluation in the population for which it is intended. High sensitivity, specificity, or AUC may demonstrate strong statistical discrimination, but clinical adoption requires evidence that the biomarker remains accurate, reproducible, feasible, and clinically useful under real-world laboratory and clinical conditions.

Challenges in Translating AI-Discovered Biomarkers into Clinical Practice

Despite the ability of AI to detect complex molecular patterns, several challenges still need to be addressed before computationally identified biomarkers can be incorporated into routine clinical diagnosis. Cancer biomarker datasets are often relatively small compared with the large number of genomic, proteomic, or metabolic features being measured, which can make models prone to overfitting and lead to inconsistent feature selection. Differences in patient characteristics, cancer stage, sample collection methods, and laboratory procedures can further affect the generalizability of these models. Variations in instruments, reagents, sample preparation, and batch processing may introduce technical differences that a model could mistakenly interpret as disease-related signals. Similarly, differences in age, sex, ethnicity, comorbidities, and disease prevalence across populations can influence model performance when a biomarker is applied in a different clinical setting. For this reason, external validation using independent and clinically representative cohorts is essential. Interpretability is another important concern, particularly when complex models are used. Feature-importance techniques and explainable AI methods such as SHAP can help identify the biomarkers that contribute to a prediction, but a computationally important feature does not necessarily represent a biologically causal biomarker. Reproducibility also depends on consistent laboratory protocols, standardized data preprocessing, clearly defined model specifications, and transparent reporting practices. Privacy and ethical concerns must also be considered when clinical and molecular datasets are combined, especially when the information could potentially be linked to individual patients. Finally, practical implementation depends on factors such as assay cost, laboratory infrastructure, computational requirements, turnaround time, regulatory assessment, and evidence that the model provides clinically meaningful information beyond existing diagnostic methods. Therefore, successful biomarker translation requires coordinated validation of the biological marker, laboratory assay, computational model, and clinical

workflow, rather than focusing solely on improving model accuracy.

Challenges in Translating AI-Discovered Biomarkers into Clinical Practice

Challenge	Clinical and technical implications
Small and heterogeneous datasets	Limited sample sizes and variation in disease stage, clinical characteristics, and specimen types can produce unstable biomarker signatures and reduce generalizability.
High dimensionality and overfitting	Thousands of molecular variables can allow a model to fit noise or cohort-specific patterns rather than robust cancer-associated signals.
Limited external validation	Strong performance within the development cohort may not persist in independent hospitals, laboratories, or populations.
Batch and laboratory variability	Differences in instruments, reagents, sample handling, and preprocessing can introduce systematic variation unrelated to disease biology.
Population differences	Age, sex, ethnicity, comorbidities, medications, disease prevalence, and cancer stage can alter biomarker distributions and predictive performance.
Model interpretability	Black-box predictions can be difficult to relate to biological mechanisms; feature importance and SHAP can improve interpretation but do not establish causality.
Reproducibility and standardization	Reliable clinical use requires standardized specimen collection, analytical procedures, preprocessing, model specifications, and reporting.
Privacy and ethics	Large molecular and clinical datasets require appropriate protection of patient information, responsible data sharing, and consideration of potential algorithmic bias.
Regulatory and clinical adoption	Translation requires evidence of analytical validity, clinical validity, clinical utility, acceptable cost, laboratory feasibility, turnaround time, and compatibility with existing diagnostic workflows.

Conclusion

Early detection remains a central objective in cancer diagnostics, and biomarkers provide an important means of identifying molecular changes associated with malignant transformation before disease becomes clinically advanced. However, conventional approaches based on individual biomarkers are often limited by insufficient sensitivity, specificity, or biological heterogeneity. Cancer does not produce a uniform molecular signal across all patients, making reliance on a single protein, nucleic acid, metabolite, or other marker inadequate for many diagnostic settings. Artificial intelligence and machine learning provide computational approaches for addressing the complexity of modern biomarker data. By analyzing high-dimensional genomic, transcriptomic, proteomic, metabolomic, and circulating biomarker profiles, these methods can identify combinations of features that may be difficult to recognize using conventional statistical

approaches. Their principal value lies not simply in increasing the number of candidate biomarkers, but in selecting informative molecular patterns and integrating complementary signals into more discriminative diagnostic signatures. Multi-biomarker and multi-omics models therefore represent an important direction for improving the molecular characterization of early cancer. Nevertheless, computational performance alone cannot establish clinical validity. AI-derived biomarkers require rigorous analytical assessment, appropriate internal validation, independent external cohorts, standardized laboratory procedures, and evaluation in populations representative of their intended clinical use. Reproducibility, interpretability, cost, assay availability, and integration with existing laboratory workflows must also be considered before implementation. Future developments are likely to increasingly combine multiple molecular and clinical data sources within multimodal AI frameworks. Integration of circulating biomarkers with

genomic, proteomic, metabolomic, imaging, and routine clinical laboratory information may provide more comprehensive representations of disease biology and improve early cancer classification. The ultimate objective is to translate these computationally identified molecular signatures into robust, reproducible, and clinically practical laboratory tests that can complement established diagnostic pathways and contribute to earlier cancer detection.

REFERENCES

1. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6), 394-424.
2. Ginsburg, O., Yip, C. H., Brooks, A., Cabanes, A., Caleffi, M., Dunstan Yataco, J. A., ... & Anderson, B. O. (2020). Breast cancer early detection: A phased approach to implementation. *Cancer*, 126, 2379-2393.
3. Mann, M., Kumar, C., Zeng, W. F., & Strauss, M. T. (2021). Artificial intelligence for proteomics and biomarker discovery. *Cell systems*, 12(8), 759-770.
4. Xiao, Q., Zhang, F., Xu, L., Yue, L., Kon, O. L., Zhu, Y., & Guo, T. (2021). High-throughput proteomics and AI for cancer biomarker discovery. *Advanced drug delivery reviews*, 176, 113844.
5. Andrades, R., & Recamonde-Mendoza, M. (2022). Machine learning methods for prediction of cancer driver genes: a survey paper. *Briefings in Bioinformatics*, 23(3), bbac062.
6. Nicora, G., Vitali, F., Dagliati, A., Geifman, N., & Bellazzi, R. (2020). Integrated multi-omics analyses in oncology: a review of machine learning methods and tools. *Frontiers in oncology*, 10, 530980.
7. Pezoulas, V. C., Hazapis, O., Lagopati, N., Exarchos, T. P., Goules, A. V., Tzioufas, A. G., ... & Gorgoulis, V. G. (2021). Machine learning approaches on high throughput NGS data to unveil mechanisms of function in biology and disease. *Cancer Genomics-Proteomics*, 18(5), 605-626.
8. Abu-Halima, M., Keller, A., Becker, L. S., Fischer, U., Engel, A., Ludwig, N., ... & Keller, V. (2023). Dynamic and static circulating cancer microRNA biomarkers—a validation study. *RNA biology*, 20(1), 1-9.
9. Abu-Halima, M., Keller, A., Becker, L. S., Fischer, U., Engel, A., Ludwig, N., ... & Keller, V. (2023). Dynamic and static circulating cancer microRNA biomarkers—a validation study. *RNA biology*, 20(1), 1-9.
10. Galal, A., Talal, M., & Moustafa, A. (2022). Applications of machine learning in metabolomics: Disease modeling and classification. *Frontiers in genetics*, 13, 1017340.
11. Nicora, G., Vitali, F., Dagliati, A., Geifman, N., & Bellazzi, R. (2020). Integrated multi-omics analyses in oncology: a review of machine learning methods and tools. *Frontiers in oncology*, 10, 530980.
12. Desaire, H., Go, E. P., & Hua, D. (2022). Advances, obstacles, and opportunities for machine learning in proteomics. *Cell Reports Physical Science*, 3(10).
13. Lee, D., Park, Y., & Kim, S. (2021). Towards multi-omics characterization of tumor heterogeneity: a comprehensive review of statistical and machine learning approaches. *Briefings in bioinformatics*, 22(3), bbaa188.
14. Pezoulas, V. C., Hazapis, O., Lagopati, N., Exarchos, T. P., Goules, A. V., Tzioufas, A. G., ... & Gorgoulis, V. G. (2021). Machine learning approaches on high throughput NGS data to unveil mechanisms of function in biology and disease. *Cancer Genomics-Proteomics*, 18(5), 605-626.
15. Schneider, L., Laiouar-Pedari, S., Kuntz, S., Krieghoff-Henning, E., Hekler, A., Kather, J. N., ... & Brinker, T. J. (2022). Integration of deep learning-based image analysis and genomic data in cancer pathology: A

- systematic review. *European journal of cancer*, 160, 80-91.
16. Kumar, S., Warrell, J., Li, S., McGillivray, P. D., Meyerson, W., Salichos, L., ... & Gerstein, M. B. (2020). Passenger mutations in more than 2,500 cancer genomes: overall molecular functional impact and consequences. *Cell*, 180(5), 915-927.
17. Cheng, F., Zhao, J., & Zhao, Z. (2016). Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. *Briefings in bioinformatics*, 17(4), 642-656.
18. Abu-Halima, M., Keller, A., Becker, L. S., Fischer, U., Engel, A., Ludwig, N., ... & Keller, V. (2023). Dynamic and static circulating cancer microRNA biomarkers—a validation study. *RNA biology*, 20(1), 1-9.
19. Li, J., Liu, K., Ji, Z., Wang, Y., Yin, T., Long, T., ... & Cheng, L. (2022). Serum Untargeted Metabolomics Analyses Reveal Metabolic Alteration of Non-Small Cell Lung Cancer and Refine Cancer Detection in Chinese Population. Available at SSRN 4153825.
20. Cardoso, M. R., Silva, A. A. R., Talarico, M. C. R., Sanches, P. H. G., Sforca, M. L., Rocco, S. A., ... & Derchain, S. F. (2022). Metabolomics by NMR combined with machine learning to predict neoadjuvant chemotherapy response for breast cancer. *Cancers*, 14(20), 5055.
21. Ye, X., Yang, X. Z., Carbone, R., Barshack, I., & Katz, R. L. (2021). Diagnosis of non-small cell lung cancer via liquid biopsy highlighting a fluorescence-in-situ-hybridization circulating tumor cell approach. In *Pathology-From Classics to Innovations*. IntechOpen.
22. Batool, S. M., Yekula, A., Khanna, P., Hsia, T., Gamblin, A. S., Ekanayake, E., ... & Balaj, L. (2023). The Liquid Biopsy Consortium: Challenges and opportunities for early cancer detection and monitoring. *Cell Reports Medicine*, 4(10).
23. Menyhárt, O., & Györffy, B. (2021). Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis.
24. Afzal, S., Hassan, M., Ullah, S., Abbas, H., Tawakkal, F., & Khan, M. A. (2022). Breast cancer; discovery of novel diagnostic biomarkers, drug resistance, and therapeutic implications. *Frontiers in molecular biosciences*, 9, 783450.
25. Xie, Y., Meng, W. Y., Li, R. Z., Wang, Y. W., Qian, X., Chan, C., ... & Leung, E. L. H. (2021). Early lung cancer diagnostic biomarker discovery by machine learning methods. *Translational oncology*, 14(1), 100907.
26. Capello, M., Bantis, L. E., Scelo, G., Zhao, Y., Li, P., Dhillon, D. S., ... & Hanash, S. M. (2017). Sequential validation of blood-based protein biomarker candidates for early-stage pancreatic cancer. *JNCI: Journal of the National Cancer Institute*, 109(4), djw266.