

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY FUTURISTIC DEVELOPMENT

Development of the OncoFusion Algorithm for Integrating Circulating Tumor DNA Methylation Mutation and Protein Biomarkers for Multi-Cancer Early Detection Compared with Conventional Single Biomarker and Ensemble Learning Models

Rukayat Olayemi Oni ^{1*}, Yewande Iyimide Adeyeye ²

¹ Department of Biology, Georgia State University, Atlanta, Georgia

² Department of Day Case Surgery, Dumfries and Galloway Royal Infirmary, Dumfries, United Kingdom

* Corresponding Author: **Rukayat Olayemi Oni**

Article Info

P-ISSN: 3051-3618

E-ISSN: 3051-3626

Volume: 03

Issue: 02

Received: 10-06-2022

Accepted: 07-07-2022

Published: 05-08-2022

Page No: 113-132

Abstract

Early detection of multiple cancer types remains a significant challenge in precision oncology because conventional diagnostic approaches frequently depend on isolated molecular biomarkers, fixed classification thresholds, and limited representations of tumour heterogeneity. Although circulating tumour DNA (ctDNA) methylation signatures, somatic mutations, and circulating protein biomarkers provide complementary diagnostic information, differences in molecular abundance, analytical sensitivity, assay reliability, and biological specificity complicate their effective integration. This study proposes OncoFusion, a novel reliability-aware, attention-guided multimodal learning algorithm designed to integrate ctDNA methylation profiles, somatic mutation features, and protein biomarker concentrations for multi-cancer early detection and cancer signal-of-origin classification. The proposed architecture combines modality-specific feature encoders, a cross-modal attention transformer, uncertainty-weighted biomarker fusion, and a multitask classification framework to distinguish cancer-positive from cancer-negative samples while identifying probable cancer origins. A reliability-adaptive fusion mechanism dynamically adjusts the contribution of each biomarker modality according to sequencing depth, molecular coverage, missingness, assay quality, and predictive uncertainty. To evaluate its effectiveness, OncoFusion is designed for comparison with conventional single-biomarker threshold classifiers, logistic regression, random forest, support vector machines, XGBoost, LightGBM, and conventional stacking ensemble models. The experimental framework incorporates stratified nested cross-validation, independent external validation, bootstrap confidence intervals, calibration analysis, and ablation experiments to establish the contribution of individual biomarkers and architectural components. Diagnostic performance will be assessed using the area under the receiver operating characteristic curve (AUROC), area under the precision–recall curve (AUPRC), sensitivity at 95% and 99% specificity, F1-score, balanced accuracy, calibration error, and cancer signal-of-origin accuracy. Comparative results will be presented through receiver operating characteristic curves, precision–recall curves, confusion matrices, calibration plots, and biomarker contribution analyses, with particular emphasis on early-stage disease and low-abundance ctDNA signals. The central hypothesis is that reliability-aware cross-modal integration can improve diagnostic discrimination, maintain high specificity, and enhance robustness under incomplete or noisy biomarker measurements compared with single-modality and conventional ensemble approaches. The proposed OncoFusion framework provides a technically testable approach to multi-cancer early detection, with potential applications in precision oncology screening and laboratory-supported diagnostic decision-making, subject to independent clinical validation.

DOI: <https://doi.org/10.54660/IJMFD.2022.3.2.113-132>

Keywords: OncoFusion Algorithm; Circulating Tumor DNA; DNA Methylation; Protein Biomarkers; Multi-Cancer Early Detection.

1. Introduction

1.1. Background of Multi-Cancer Early Detection and Precision Oncology Diagnostics

Multi-cancer early detection (MCED) represents an important development in precision oncology, seeking to identify molecular evidence of malignancy across different anatomical sites before advanced clinical manifestations become apparent. Conventional

cancer-screening programmes generally target individual malignancies through organ-specific procedures, including mammography, colonoscopy, cervical cytology, and low-dose computed tomography. Although these approaches provide established clinical benefits in appropriate populations, several aggressive cancers lack effective population-level screening strategies. Advances in liquid biopsy have introduced opportunities to interrogate tumour-derived molecular information using peripheral blood, potentially expanding screening coverage beyond individual organs. Aravanis *et al.* (2017) identified next-generation sequencing of circulating tumour DNA as a promising foundation for early cancer detection, particularly through highly sensitive identification of tumour-associated genomic abnormalities. Nevertheless, the extremely low abundance of tumour-derived molecules during early disease requires analytical systems capable of distinguishing biological signals from sequencing errors, normal cellular variation, and background molecular noise. Consequently, effective MCED requires more than highly sensitive laboratory measurements; it also requires computational methods that integrate heterogeneous biomarkers while preserving diagnostic specificity.

The emergence of precision oncology diagnostics has strengthened interest in combining molecular features representing different dimensions of cancer biology. DNA methylation patterns provide information about epigenetic dysregulation and tissue identity, somatic mutations characterize tumour-associated genomic alterations, and circulating proteins reflect downstream biological processes involving tumour growth, inflammation, and host responses. Lo and Lam (2020) discussed the potential of blood-based multi-cancer screening using circulating DNA and protein biomarkers, highlighting the clinical relevance of combining molecular testing with confirmatory diagnostic procedures. Building on this foundation, the proposed OncoFusion algorithm establishes a unified computational architecture for integrating methylation, mutation, and protein measurements. Its principal design objective is to identify complementary molecular evidence through modality-specific encoders, cross-modal attention, and reliability-adaptive fusion. Unlike fixed biomarker combinations, OncoFusion is designed to account for sequencing depth, assay quality, biomarker missingness, and predictive uncertainty when estimating cancer probability and probable cancer origin. The intended result is a clinically interpretable diagnostic framework whose effectiveness can be established through comparative testing against conventional single-biomarker classifiers and ensemble learning models.

1.2. Biological Significance of Circulating Tumor DNA in Early Cancer Detection

Circulating tumour DNA constitutes a tumour-derived fraction of circulating cell-free DNA released into biological fluids through cellular apoptosis, necrosis, and other mechanisms associated with tumour activity. Its molecular composition can contain somatic substitutions, insertions, deletions, copy-number alterations, and epigenetic abnormalities originating from malignant cell populations. Unlike tissue biopsy, which samples a restricted anatomical

region, plasma-derived ctDNA may capture molecular information released from spatially distinct tumour subpopulations. However, its diagnostic accessibility depends strongly on tumour volume, vascularization, anatomical location, biological turnover, and the efficiency of DNA release into circulation. Bettegowda *et al.* (2014) demonstrated substantial variation in ctDNA detectability across cancer types and clinical stages, establishing that early-stage malignancies frequently generate considerably weaker circulating signals than advanced cancers. This finding has direct implications for OncoFusion because mutation-negative plasma samples cannot automatically be interpreted as cancer-free. Instead, the algorithm must distinguish true biological absence from inadequate molecular sampling, insufficient sequencing coverage, and low variant allele fractions, particularly when tumour-derived fragments approach the analytical detection limit.

Tumour evolution further complicates ctDNA-based diagnosis because malignant populations accumulate genetically heterogeneous alterations through clonal expansion, selection, and treatment-independent evolutionary processes. Abbosh *et al.* (2017) demonstrated that phylogenetic analysis of circulating DNA could reveal clonal and subclonal features of early-stage non-small-cell lung cancer, establishing the importance of genomic architecture in interpreting plasma-derived mutation signals. For multi-cancer classification, this evidence suggests that mutation features should include variant allele frequency, mutation burden, clonal representation, and the biological relevance of alterations rather than relying exclusively on binary mutation detection. Accordingly, OncoFusion is designed to encode mutation-derived features within a dedicated genomic representation that preserves information about variant abundance and reliability. The encoded mutation representation is subsequently integrated with methylation signatures and protein biomarkers through attention-based fusion. This structure is intended to compensate for incomplete mutation information by exploiting complementary epigenetic and proteomic evidence without treating missing signals as definitive negatives. Evaluation should therefore examine diagnostic sensitivity across tumour stages, variant allele frequency intervals, and sequencing coverage categories to establish whether multimodal integration improves early-stage discrimination under biologically realistic conditions.

1.3. DNA Methylation Signatures and Epigenetic Alterations as Cancer-Specific Biomarkers

DNA methylation constitutes a fundamental epigenetic mechanism regulating transcriptional activity, chromatin organization, cellular differentiation, and genomic stability. In cancer, abnormal methylation frequently involves promoter-associated hypermethylation, widespread hypomethylation, and disruption of tissue-specific regulatory patterns. These alterations can emerge during malignant transformation and may provide diagnostic information even when individual somatic mutations occur infrequently. Methylation signatures are particularly relevant to MCED because their genomic distributions can reflect both malignant transformation and the tissue from which

circulating DNA originated. Shen *et al.* (2018) demonstrated that plasma cell-free DNA methylome profiling could support tumour detection and classification across several malignancies, providing evidence that epigenetic information offers discriminatory characteristics beyond conventional mutation sequencing. From a computational perspective, methylation measurements can be represented through differentially methylated regions, CpG-level methylation fractions, regional methylation densities, and tumour-associated epigenetic patterns. Nevertheless, background methylation from normal blood cells, variation in sequencing coverage, and technical differences in methylation enrichment or conversion procedures introduce uncertainty that must be addressed before reliable classification.

The diagnostic potential of methylation profiling was further established by Liu *et al.* (2020), who investigated targeted methylation signatures for detecting and localizing more than 50 cancer types using circulating cell-free DNA. Their findings support the feasibility of combining high-specificity cancer detection with cancer signal-of-origin prediction through appropriately selected epigenetic features. However, diagnostic sensitivity remains influenced by cancer stage, molecular abundance, and differences between tumour types. OncoFusion therefore incorporates methylation as a primary molecular modality within a broader multimodal learning system rather than treating it as an independently sufficient diagnostic indicator. Its methylation encoder is designed to learn nonlinear relationships among cancer-associated genomic regions while accounting for feature sparsity, methylation measurement confidence, and sample-specific coverage. The resulting latent representation will interact with mutation and protein embeddings through cross-modal attention, allowing the algorithm to learn whether epigenetic abnormalities are reinforced by independent molecular evidence. Comparative experiments should quantify the incremental contribution of methylation using modality-specific ablation, tissue-of-origin classification accuracy, and sensitivity at predetermined specificity thresholds. Such evaluation is necessary to determine whether reliability-adaptive fusion offers measurable advantages over methylation-only classifiers and conventional feature-concatenation approaches.

1.4. Somatic Mutation Profiles and Genomic Heterogeneity in Circulating Tumor DNA

Somatic mutations provide important molecular evidence of malignant transformation because cancer development involves the progressive accumulation of genomic alterations affecting cellular proliferation, apoptosis, DNA repair, and signalling pathways. Circulating tumour DNA can contain substitutions, insertions, deletions, copy-number abnormalities, and structural alterations originating from genetically heterogeneous tumour populations. Frequently investigated oncogenic alterations include mutations affecting KRAS, TP53, EGFR, BRAF, and PIK3CA, although their prevalence and diagnostic relevance differ substantially across malignancies. The interpretation of these alterations becomes particularly challenging during early-stage disease because tumour-derived DNA may represent only a small proportion of total circulating cell-free DNA.

Furthermore, mutations detected in plasma do not necessarily originate from malignant solid tumours. Razavi *et al.* (2019) demonstrated that high-intensity sequencing can identify plasma variants arising from different biological sources, including clonal haematopoiesis, thereby emphasizing the importance of distinguishing tumour-derived mutations from background alterations. These findings establish a critical requirement for the proposed OncoFusion algorithm: mutation classification must incorporate variant-specific biological context, molecular abundance, sequencing confidence, and potential non-tumour origins rather than assuming that every detected alteration represents malignancy.

Genomic heterogeneity introduces additional difficulties because different tumour subclones may possess distinct mutational profiles, resulting in variable biomarker detectability across patients with the same cancer type. Chabon *et al.* (2020) demonstrated that integrating genomic features could improve non-invasive early lung cancer detection, illustrating the diagnostic importance of combining mutation-related information with broader molecular characteristics. Within OncoFusion, somatic mutations are therefore represented through structured features incorporating variant allele frequency, gene identity, functional annotation, mutation burden, and sequencing reliability. A modality-specific encoder is designed to transform these features into a latent genomic representation while reducing sensitivity to sparse mutation observations and technical artifacts. The proposed cross-modal attention mechanism subsequently evaluates whether mutation-associated evidence is consistent with methylation-derived tissue signatures and circulating protein patterns. For example, a low-frequency KRAS alteration accompanied by compatible epigenetic and proteomic features may receive greater diagnostic consideration than an isolated mutation lacking independent molecular support. However, such integration must be trained and validated against appropriate controls to prevent spurious associations. The resulting comparative analysis will examine whether OncoFusion improves mutation classification robustness, early-stage sensitivity, and specificity relative to mutation-only classifiers and conventional ensemble models.

1.5. Circulating Protein Biomarkers and Their Complementary Diagnostic Contributions

Circulating protein biomarkers provide complementary information about tumour biology because malignant transformation alters protein expression, secretion, cellular signalling, inflammatory activity, and interactions between tumour cells and their surrounding microenvironment. Unlike ctDNA mutations and methylation signatures, which primarily represent genomic and epigenomic abnormalities, circulating proteins reflect functional consequences of biological processes occurring within malignant tissues and host systems. Commonly investigated cancer-associated proteins include carcinoembryonic antigen, cancer antigen 125, carbohydrate antigen 19-9, alpha-fetoprotein, and prostate-specific antigen. Nevertheless, individual proteins frequently demonstrate inadequate specificity because their concentrations may increase during benign inflammatory,

metabolic, or physiological conditions. Borrebaeck (2017) emphasized the importance of progressing from isolated protein measurements toward multiplex biomarker signatures capable of improving diagnostic discrimination. This perspective is particularly relevant to MCED, where biomarker panels must identify diverse malignancies without generating excessive false-positive classifications. Consequently, protein features should be evaluated through normalized concentration measurements, analytical detection limits, biological variability, inter-assay precision, and associations with non-malignant conditions. These requirements establish the foundation for incorporating proteomic reliability indicators into the OncoFusion architecture.

The diagnostic value of combining protein measurements with genomic information was demonstrated by Cohen *et al.* (2018), who developed CancerSEEK, a multi-analyte blood test integrating circulating protein biomarkers with cell-free DNA mutations. Their study evaluated 1,005 patients with nonmetastatic cancers across eight tumour types and reported a median detection sensitivity of approximately 70%, with specificity exceeding 99% among the evaluated healthy controls. Although these findings support multimodal detection, the performance of fixed biomarker combinations may vary across cancer types, clinical stages, and assay conditions. OncoFusion extends the underlying integration concept by incorporating DNA methylation as a third molecular modality and introducing reliability-adaptive weighting to regulate the influence of individual biomarker groups. Its protein encoder will capture nonlinear concentration relationships, while cross-modal attention will evaluate interactions between proteomic patterns and tumour-associated genomic abnormalities. The proposed framework will also investigate whether protein measurements improve detection when ctDNA mutation abundance is insufficient for reliable classification. Ablation experiments comparing methylation-only, mutation-only, protein-only, pairwise fusion, and complete three-modality models will establish the incremental diagnostic contribution of circulating proteins under controlled validation conditions.

1.6. Limitations of Conventional Single-Biomarker Detection and Ensemble Learning Models

Conventional single-biomarker cancer detection methods commonly depend on predefined molecular thresholds, individual mutation signatures, isolated methylation measurements, or circulating protein concentrations. Although these approaches can provide clinically useful information, their performance is constrained by biological heterogeneity, low biomarker abundance, and overlap between malignant and non-malignant molecular distributions. Fixed thresholds are particularly problematic when diagnostic signals vary with cancer stage, tumour origin, specimen quality, and analytical platform. Cristiano *et al.* (2019) demonstrated that genome-wide cell-free DNA fragmentation patterns contain cancer-associated information beyond individual genomic alterations, supporting the broader principle that diagnostically relevant molecular evidence is distributed across multiple biological characteristics. Conventional machine learning classifiers,

including logistic regression, support vector machines, random forests, and gradient-boosted decision trees, can model complex biomarker relationships, but their effectiveness depends on feature representation, training population characteristics, and validation design. Similarly, conventional stacking and voting ensembles combine classifier outputs without necessarily accounting explicitly for assay reliability or sample-specific molecular uncertainty. When modalities differ substantially in completeness or measurement quality, indiscriminate feature concatenation or fixed-weight prediction aggregation may produce unstable classifications and poorly calibrated probabilities.

The limitations of single-modality approaches become particularly apparent when diagnostic performance is evaluated across clinical stages rather than through aggregate discrimination alone. Klein *et al.* (2021) independently validated a targeted methylation-based MCED test and reported overall sensitivity of 51.5%, with stage I sensitivity of 16.8% and substantially higher sensitivity in advanced disease. These findings demonstrate that high-specificity molecular classification does not necessarily translate into adequate detection of early malignancies. OncoFusion addresses this methodological challenge through a proposed reliability-aware fusion mechanism that assigns sample-dependent contributions to methylation, mutation, and protein representations. Its cross-modal attention architecture is designed to identify complementary evidence, while uncertainty estimation will distinguish confident predictions from classifications affected by insufficient molecular information. Comparative evaluation against logistic regression, random forest, support vector machines, XGBoost, LightGBM, and stacking ensembles will assess whether these mechanisms provide genuine improvements. Performance will be examined using AUROC, AUPRC, sensitivity at 95% and 99% specificity, calibration error, and cancer signal-of-origin accuracy. Independent validation and ablation analysis will be essential for determining whether any observed advantage results from meaningful multimodal integration rather than increased model complexity.

1.7. Research Gap and Motivation for Reliability-Aware Multimodal Biomarker Fusion

Despite advances in liquid biopsy technologies and artificial intelligence-based cancer classification, important methodological challenges remain in the integration of heterogeneous molecular biomarkers for early-stage multi-cancer detection. Existing approaches have demonstrated the diagnostic value of mutation profiling, methylation analysis, circulating proteins, and selected multimodal combinations; however, analytical uncertainty and differences in biomarker detectability remain major barriers to dependable clinical application. Pantel and Alix-Panabières (2019) identified low concentrations of circulating tumour-derived material as a fundamental challenge for highly sensitive liquid biopsy applications, particularly when residual or early malignant disease produces limited molecular evidence. This problem extends beyond laboratory sensitivity because computational classifiers may interpret absent or unreliable measurements as biologically meaningful negative observations. Furthermore, different biomarker modalities exhibit distinct

measurement distributions, technical error structures, and relationships with cancer stage. A central research gap therefore concerns the development and rigorous evaluation of a unified model that can jointly learn methylation, mutation, and protein representations while explicitly adjusting their contributions according to sample-specific analytical reliability. The proposed OncoFusion framework is motivated by this requirement for biologically informed, uncertainty-sensitive multimodal learning rather than conventional fixed-weight biomarker aggregation. Clinical implementation introduces additional challenges involving assay standardization, population heterogeneity, reproducibility, diagnostic calibration, and the interpretation of molecular findings. Ignatiadis *et al.* (2021) emphasized that the translation of liquid biopsy into clinical practice requires careful consideration of analytical validity, clinical validity, and implementation constraints. In response, OncoFusion proposes a modality-specific encoding architecture integrated through cross-modal attention, reliability-dependent weighting, and multitask prediction of cancer presence and probable cancer origin. The model is designed to incorporate sequencing depth, methylation coverage, protein assay precision, missingness indicators, and predictive uncertainty as explicit computational inputs. Its novelty will be evaluated against established single-biomarker classifiers and conventional ensemble learning approaches using identical datasets, consistent preprocessing, and independent validation procedures. Particular attention will be directed toward sensitivity in stage I and II cancers, false-positive control, calibration stability, and robustness under incomplete biomarker availability. The study therefore seeks to determine whether reliability-aware multimodal integration provides a measurable diagnostic advantage while maintaining clinically relevant specificity. Any superiority claim will depend on reproducible experimental evidence, statistically supported comparisons, and validation across independent patient populations.

1.8. Research Objectives and Research Questions

The primary objective of this study is to develop and evaluate OncoFusion, a reliability-aware multimodal artificial intelligence algorithm that integrates ctDNA methylation signatures, somatic mutation profiles, and circulating protein biomarkers for multi-cancer early detection and cancer signal-of-origin prediction. The investigation is structured around five research objectives and five corresponding research questions to establish the algorithm's diagnostic effectiveness, computational reliability, and comparative advantages over conventional single-biomarker and ensemble learning models.

• Research Objectives

1. To develop a novel OncoFusion architecture integrating ctDNA methylation, somatic mutation, and circulating protein biomarker representations through modality-specific encoders and cross-modal attention mechanisms.
2. To formulate a reliability-adaptive biomarker fusion mechanism incorporating sequencing depth, methylation coverage, protein assay precision, missing biomarker information, and predictive uncertainty.

3. To evaluate OncoFusion's diagnostic performance using AUROC, AUPRC, sensitivity, specificity, F1-score, calibration error, and cancer signal-of-origin classification accuracy.
4. To compare OncoFusion against conventional single-biomarker classifiers, logistic regression, support vector machines, random forests, XGBoost, LightGBM, and stacking ensemble models under consistent experimental conditions.
5. To investigate the contributions of individual biomarker modalities and fusion components through ablation analysis, early-stage cancer evaluation, robustness testing, and independent external validation.

• Research Questions

1. How effectively can OncoFusion integrate ctDNA methylation, somatic mutation, and circulating protein biomarkers to distinguish cancer-positive from cancer-negative samples?
2. To what extent does reliability-adaptive weighting improve diagnostic robustness when molecular measurements exhibit low abundance, analytical noise, or incomplete biomarker coverage?
3. How accurately can OncoFusion detect early-stage malignancies and predict cancer signal-of-origin while maintaining clinically appropriate specificity?
4. Does OncoFusion achieve statistically significant improvements in diagnostic discrimination and calibration compared with conventional single-biomarker classifiers and established ensemble learning algorithms?
5. What are the relative contributions of methylation, mutation, protein biomarkers, cross-modal attention, and reliability weighting to overall diagnostic performance and external generalizability?

1.9. Novel Contributions and Significance of the Study

This study proposes OncoFusion as a unified computational framework for multi-cancer early detection through the integration of complementary genomic, epigenomic, and proteomic biomarkers. Its principal methodological contribution is the development of a reliability-aware cross-modal attention architecture that combines modality-specific feature learning with sample-dependent biomarker weighting, allowing molecular evidence to be evaluated according to both biological relevance and analytical quality. Unlike conventional feature-concatenation approaches and fixed-weight ensemble classifiers, OncoFusion is designed to incorporate sequencing coverage, variant allele frequency, methylation measurement reliability, protein assay variability, missing biomarker information, and predictive uncertainty into the fusion process. A further contribution is the formulation of a multitask classification framework capable of estimating cancer probability and predicting cancer signal-of-origin through shared molecular representations. The study also proposes a systematic experimental strategy involving single-modality baselines, conventional machine learning algorithms, ensemble classifiers, ablation experiments, probability calibration, and independent validation. Its scientific significance lies in testing whether three-modality integration provides

additional diagnostic information beyond existing single- and dual-modality approaches, particularly for early-stage malignancies characterized by limited ctDNA abundance. Its potential clinical significance concerns improved diagnostic confidence, reduced susceptibility to unreliable biomarker measurements, and more informative molecular classification. These contributions remain proposed innovations whose effectiveness and originality must be established through implementation, comparison with the existing literature, and reproducible validation.

1.10. Scope of the Review and Structure of the Paper

This study focuses on the computational development, mathematical formulation, and comparative evaluation of the proposed OncoFusion algorithm for multi-cancer early detection using circulating tumour DNA methylation profiles, somatic mutation features, and circulating protein biomarkers. Its methodological scope includes biomarker preprocessing, feature normalization, modality-specific representation learning, reliability estimation, cross-modal attention, uncertainty-aware fusion, multitask classification, and comparative machine learning evaluation. The experimental design will prioritize datasets containing matched biomarker measurements and independently established cancer diagnoses, with particular consideration given to cancer stage, tissue of origin, assay quality, and non-malignant control populations. Diagnostic performance will be investigated through discrimination metrics, sensitivity at fixed specificity, calibration assessment, ablation studies, and external validation. The study does not include the development of new wet-laboratory assays, prospective patient recruitment, or claims of demonstrated clinical utility without supporting evidence. The manuscript is organized into five principal sections: Section 1 establishes the background, biological foundations, research problem, objectives, and proposed contributions; Section 2 critically reviews existing liquid biopsy technologies, biomarker integration strategies, and conventional diagnostic algorithms; Section 3 presents the mathematical formulation, architecture, optimization procedures, and experimental design of OncoFusion; Section 4 examines comparative diagnostic performance, early-stage detection, cancer signal-of-origin classification, biomarker contributions, and robustness; and Section 5 presents the research findings, methodological limitations, recommendations for clinical validation, and directions for further algorithm development.

2. Literature Review

2.1. Circulating Tumor DNA Biology, Liquid Biopsy Technologies, and Molecular Detection Challenges

Circulating tumour DNA (ctDNA) constitutes a biologically

informative fraction of circulating cell-free DNA originating from malignant tissues through apoptosis, necrosis, and other DNA-release mechanisms. Its molecular characteristics include tumour-associated mutations, epigenetic alterations, and fragment-length distributions that can support minimally invasive cancer detection. Wan *et al.* (2017) established the relevance of ctDNA analysis for molecular oncology while identifying analytical sensitivity, limited tumour-derived DNA abundance, and preanalytical variability as substantial barriers to clinical implementation as represented in figure 1. Similarly, Heitzer *et al.* (2019) examined the expanding role of liquid biopsies in genomics-driven oncology, emphasizing the importance of assay standardization and reliable interpretation of low-frequency molecular signals. Early-stage tumours present particularly demanding detection conditions because ctDNA concentrations may approach the limits of molecular measurement technologies. Consequently, sequencing depth, molecular barcoding, error suppression, sample processing, and variant-calling thresholds influence diagnostic reliability. These challenges motivate the proposed OncoFusion framework, which integrates mutation, methylation, and protein information rather than interpreting individual molecular observations independently.

The computational management of multimodal liquid-biopsy information introduces additional challenges involving secure data exchange, institutional heterogeneity, and reproducible processing. Atalor (2019), in the supplied citation, addresses privacy-preserving federated learning for oncology-related adverse drug event prediction; the concept is potentially transferable to distributed OncoFusion training, although its reported application is not ctDNA detection. Likewise, the supplied work of Nwokocha *et al.* (2021a) concerns FHIR-based interoperability, a relevant consideration for integrating laboratory results and clinical metadata across healthcare systems. Their separate study on ECMS integration and Smile CDR adoption (Nwokocha *et al.*, 2021b) provides a related systems-integration context rather than molecular diagnostic evidence. Together, these methodological considerations support an OncoFusion implementation in which molecular data undergo standardized preprocessing, quality assessment, and secure integration before classification. The proposed model should distinguish biological uncertainty from technical missingness and incorporate assay-specific reliability indicators into its multimodal representation. This design is particularly important when mutation detection is weak but methylation or protein measurements provide complementary evidence of malignancy.

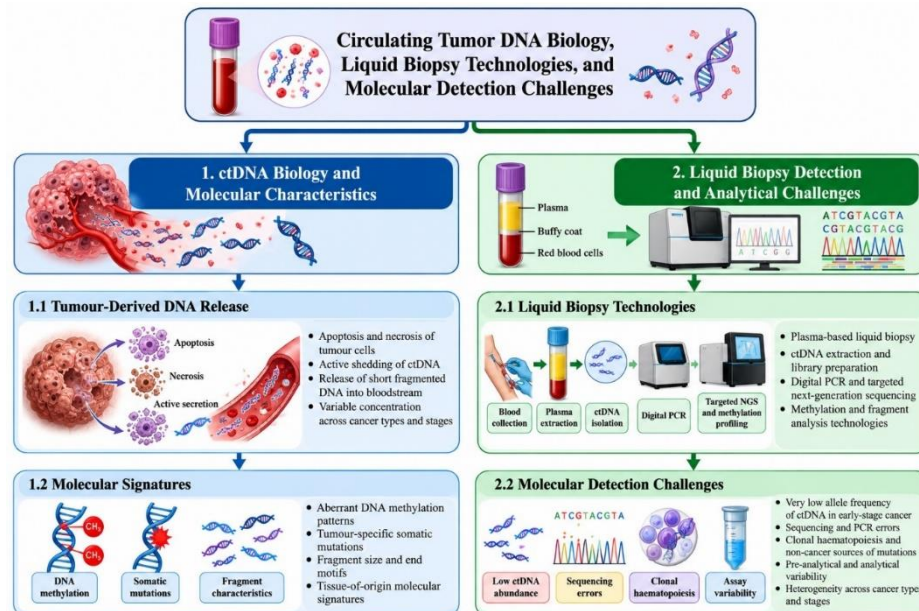


Fig 1: Biological origins, molecular signatures, liquid biopsy technologies, and analytical challenges of circulating tumour DNA detection.

Figure 1 illustrates the biological origins, molecular characteristics, liquid biopsy technologies, and analytical challenges associated with circulating tumour DNA (ctDNA) detection, establishing the molecular and computational foundation of the proposed OncoFusion algorithm for multi-cancer early detection. The diagram is organized into two interconnected branches. The first branch describes ctDNA biology and molecular characteristics, demonstrating how malignant cells release tumour-derived DNA fragments into peripheral circulation through apoptosis, necrosis, and other DNA-shedding mechanisms. These fragments contain diagnostically relevant molecular signatures, including aberrant CpG methylation patterns, tumour-associated somatic mutations, and fragmentomic characteristics such as fragment length distributions and end motifs. DNA methylation signatures provide information about epigenetic dysregulation and tissue of origin, whereas somatic mutations reveal oncogenic alterations, clonal diversity, and tumour-associated genomic abnormalities. The second branch illustrates the liquid biopsy workflow, beginning with peripheral blood collection, plasma separation, and cell-free DNA extraction, followed by molecular analysis using digital polymerase chain reaction (dPCR), targeted next-generation sequencing (NGS), and methylation profiling technologies. These methods support the identification and quantification of low-abundance tumour-derived molecular signals, although their diagnostic effectiveness depends on sequencing coverage, analytical sensitivity, sample integrity, and molecular error suppression. The diagram further identifies major detection challenges, including extremely low ctDNA concentrations in early-stage malignancies, sequencing and amplification errors, clonal haematopoiesis-associated mutations, preanalytical variability, and differences in tumour shedding across cancer types and clinical stages. Within the OncoFusion framework, these challenges motivate the incorporation of modality-specific feature encoding, assay reliability modelling, and uncertainty-adaptive cross-modal attention to distinguish biologically informative signals from technical noise. Although the diagram concentrates on ctDNA-based detection, the complete OncoFusion architecture additionally

integrates circulating protein biomarkers to provide complementary diagnostic evidence, particularly when mutation or methylation measurements are incomplete or unreliable.

2.2. DNA Methylation, Somatic Mutation, and Protein Biomarker Signatures for Multi-Cancer Classification

Multi-cancer classification requires integration of molecular characteristics that capture complementary dimensions of malignant transformation, tissue identity, and systemic biological responses. DNA methylation signatures characterize epigenetic abnormalities through altered CpG methylation patterns, whereas somatic mutations reveal genomic changes associated with oncogenic transformation and clonal evolution. Circulating protein biomarkers provide additional information about tumour-associated secretion, metabolic activity, inflammatory signalling, and host responses. Sun *et al.* (2015) demonstrated that genome-wide methylation analysis of plasma DNA could identify tissue contributions to circulating DNA, establishing an important biological foundation for cancer signal-of-origin classification as presented in table 1. Best *et al.* (2015) independently demonstrated the diagnostic potential of tumour-educated platelet RNA profiles for pan-cancer and multiclass prediction, reinforcing the broader value of molecular information beyond DNA mutations. Although platelet transcriptomics is distinct from circulating protein quantification, these findings support the biological rationale for integrating complementary blood-derived modalities. Accordingly, OncoFusion proposes separate encoders for methylation-region measurements, somatic mutation profiles, and normalized protein concentrations, enabling each molecular domain to retain its distinctive statistical characteristics before fusion.

The methodological challenge lies in integrating heterogeneous features without allowing measurement scale, missingness, or technical variability to distort diagnostic inference. The supplied work of Onyekiaonwu *et al.* (2019) concerns predictive modelling in community pharmacy applications and offers a general machine learning perspective rather than direct evidence for molecular cancer

classification. Similarly, Oladoye *et al.* (2021), as cited in the supplied bibliography, address sensor fusion and degradation modelling, providing a conceptual analogy for combining measurements with differing reliability. Aluso (2021) discusses cross-platform data integration in marketing analytics, which is relevant only to the broader problem of harmonizing heterogeneous information. OncoFusion translates these general computational principles into a biologically specific architecture using modality-wise normalization, quality indicators, learned feature

embeddings, and attention-based interactions. For example, weak mutation evidence accompanied by highly informative methylation patterns and consistent protein concentrations may support classification more effectively than mutation detection alone. The model will therefore be evaluated using single-modality, pairwise-modality, and complete three-modality experiments to quantify complementary information and determine whether reliability-aware integration improves early-stage sensitivity without compromising specificity.

Table 1. Comparative Characteristics of DNA Methylation, Somatic Mutation, and Protein Biomarker Signatures for Multi-Cancer Classification

Biomarker Modality	Molecular Characteristics and Diagnostic Signatures	Contribution to Multi-Cancer Classification	Integration Within the Proposed OncoFusion Algorithm
DNA Methylation	Aberrant CpG methylation, promoter hypermethylation, genome-wide hypomethylation, and tissue-specific epigenetic signatures.	Supports cancer detection, tissue-of-origin identification, and discrimination between malignant and non-malignant molecular profiles.	A methylation-specific encoder extracts epigenetic representations, while cross-modal attention incorporates sequencing coverage and methylation measurement reliability.
Somatic Mutations	Tumour-associated alterations in genes such as TP53, KRAS, EGFR, BRAF, and PIK3CA, characterized by variant allele frequency and mutation patterns.	Identifies oncogenic alterations, supports molecular characterization, and provides complementary evidence for distinguishing cancer types.	A genomic encoder processes mutation features, sequencing confidence, and variant abundance, enabling reliability-weighted integration with methylation and protein representations.
Protein Biomarkers	Circulating concentrations of proteins such as CEA, CA-125, CA19-9, AFP, and PSA, reflecting tumour-associated secretion and host biological responses.	Provides complementary biochemical evidence that may improve cancer discrimination when ctDNA signals are weak or insufficient.	A protein-specific encoder captures concentration-dependent patterns, while uncertainty-weighted fusion adjusts contributions according to assay precision and biomarker reliability.

2.3. Conventional Threshold-Based Methods, Logistic Regression, and Single-Biomarker Diagnostic Algorithms

Conventional biomarker-based diagnostic algorithms typically classify samples through predetermined concentration thresholds, mutation-presence indicators, or statistical relationships between measured features and disease status. Threshold-based methods are computationally straightforward and clinically interpretable, but their effectiveness depends on the separation between cancer-positive and cancer-negative biomarker distributions. Logistic regression extends this approach by estimating cancer probability from weighted combinations of diagnostic predictors, allowing continuous biomarker measurements and selected clinical covariates to contribute to classification. However, linear decision functions may inadequately represent complex interactions among methylation, mutation, and protein measurements as represented in figure 2. Pepe *et al.* (2014) examined the limitations of risk reclassification statistics, demonstrating the importance of rigorous assessment when evaluating improvements in predictive models. Diagnostic comparison should therefore extend beyond apparent accuracy to include discrimination, calibration, and clinically relevant operating thresholds. Robin *et al.* (2011) developed the pROC framework for receiver operating characteristic analysis and statistical comparison, providing methodological support for evaluating classifiers across sensitivity–specificity trade-offs. Although that publication predates the requested period, it is retained here as a foundational statistical reference rather than misrepresented as a 2013–2021 publication. Decision thresholds also have important conceptual parallels with classification rules used outside biomedical diagnostics, although such parallels do not establish clinical validity. Ajayi *et al.* (2019) address legal prosecutability and

evidentiary standards rather than statistical cancer diagnosis; consequently, their work cannot substantiate molecular detection performance. Kwarteng *et al.* (2020) examine cybersecurity risk reduction through access-control mechanisms, offering only a general analogy for rule-based decisions and false-alert management. Amebleh (2021) discusses survival and hierarchical Bayesian forecasting in financial liability modelling, illustrating a transferable statistical modelling concept rather than an oncology application. For OncoFusion, the relevant methodological issue is whether a fixed diagnostic threshold or conventional logistic regression can adequately classify samples when molecular measurements differ in abundance, quality, and biological specificity. The proposed comparison will therefore include mutation-only, methylation-only, and protein-only threshold models alongside regularized logistic regression and other conventional classifiers. All models should be assessed using identical training and testing partitions, with operating thresholds selected exclusively from training data. This strategy will determine whether OncoFusion's nonlinear multimodal representations provide improvements beyond simpler, interpretable diagnostic baselines. Figure 2. illustrates the computational principles, diagnostic applications, and methodological limitations of conventional threshold-based methods, logistic regression, and single-biomarker machine learning algorithms used in multi-cancer early detection. The diagram is organized into three interconnected branches representing distinct approaches to molecular cancer classification. The first branch describes conventional threshold-based diagnostic methods, in which cancer-positive and cancer-negative samples are distinguished using predefined cutoffs for circulating tumour DNA (ctDNA) variant allele frequency, CpG methylation

levels, or circulating protein concentrations, including carcinoembryonic antigen (CEA), cancer antigen 125 (CA-125), and prostate-specific antigen (PSA). Although these methods offer computational simplicity and straightforward clinical interpretation, fixed thresholds may produce false-negative results when tumour-derived molecular signals fall below analytical detection limits, particularly in stage I and II malignancies, while benign conditions and clonal haematopoiesis may contribute to false-positive classifications. The second branch presents logistic regression and statistical classification models, which transform normalized biomarker measurements into estimated cancer probabilities using a weighted linear predictor and sigmoid activation function, expressed as $P(Y = 1|X) = 1/[1 + \exp(-(\beta_0 + \sum_{j=1}^p \beta_j X_j))]$, where β_0 represents the intercept, β_j denotes the regression coefficient for biomarker j , and X_j represents its measured value. Despite their interpretability, conventional logistic regression models may inadequately represent nonlinear molecular interactions unless appropriate interaction terms or transformations are explicitly incorporated. The third branch examines single-biomarker machine learning algorithms, including support

vector machines, random forests, and gradient-boosted classifiers trained separately on methylation signatures, somatic mutation characteristics, or protein biomarker profiles. These algorithms can identify complex patterns within individual molecular datasets but cannot exploit complementary genomic, epigenomic, and proteomic information when restricted to one modality. Furthermore, inadequate treatment of sequencing errors, missing biomarker measurements, assay variability, and predictive uncertainty may reduce diagnostic reliability and external generalizability. Collectively, the diagram establishes the methodological motivation for the proposed OncoFusion algorithm, which combines modality-specific feature encoders, reliability-adaptive cross-modal attention, uncertainty-weighted fusion, and multitask classification to investigate whether integrated molecular evidence can improve early-stage cancer detection, cancer signal-of-origin prediction, probability calibration, and sensitivity at clinically relevant specificity thresholds compared with conventional single-biomarker and ensemble learning approaches.

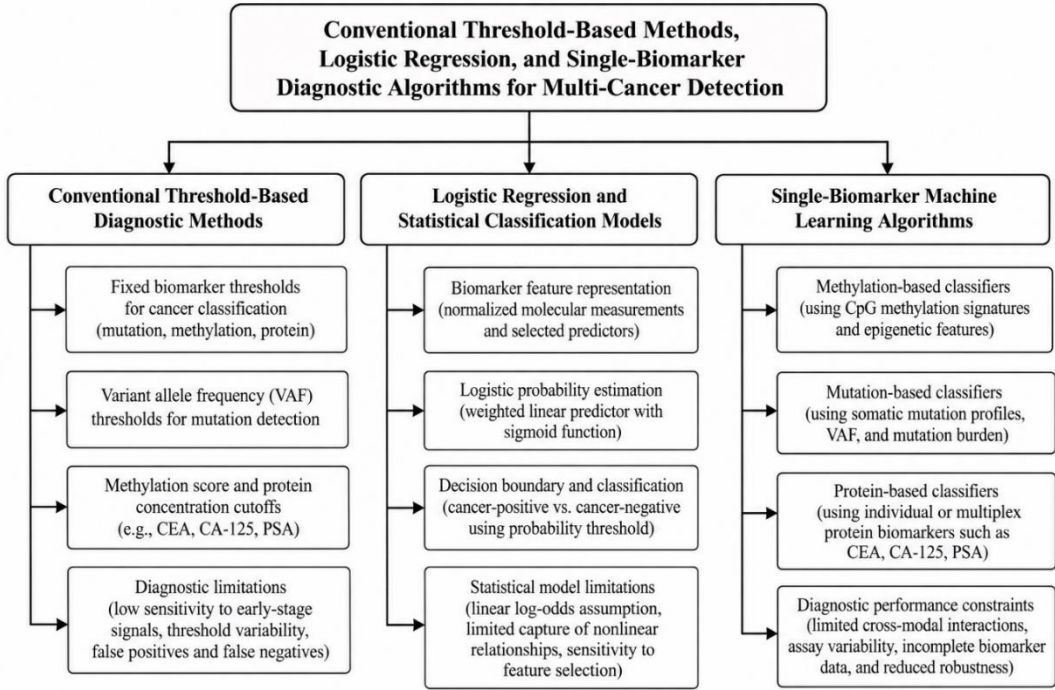


Fig 2: Conceptual classification of threshold-based, logistic regression, and single-biomarker machine learning approaches for multi-cancer detection.

2.4. Machine Learning, Ensemble Learning, Deep Learning, and Transformer-Based Multi-Omics Integration

Machine learning and deep learning have expanded the computational capacity of molecular diagnostics by enabling nonlinear representation learning, high-dimensional feature selection, and integration of heterogeneous biological measurements. Conventional algorithms such as support vector machines, random forests, and gradient-boosted decision trees provide useful diagnostic baselines, whereas deep neural networks can learn hierarchical relationships among complex molecular variables as shown in figure 3. Argelaguet *et al.* (2018) introduced Multi-Omics Factor Analysis, demonstrating how shared latent factors can

represent structured variation across heterogeneous omics datasets. Although primarily an unsupervised framework, this approach establishes an important foundation for learning relationships between molecular modalities. Ching *et al.* (2018) examined opportunities and limitations of deep learning in biomedical applications, including interpretability, limited sample sizes, data heterogeneity, and overfitting. These challenges are especially relevant to multi-cancer detection because methylation profiles may contain thousands of genomic features, somatic mutations are frequently sparse, and protein measurements exhibit distinct distributions. OncoFusion addresses these differences through modality-specific encoders followed by a cross-modal attention mechanism designed to learn interactions

between epigenetic, genomic, and proteomic representations while incorporating assay reliability into the fusion process. The supplied literature also contains computational studies outside oncology whose relevance should be interpreted carefully. Idika *et al.* (2021) describe deep learning for malware classification, providing a domain-independent example of automated feature learning but no direct evidence of diagnostic performance in cancer populations. Ijiga *et al.* (2021a) investigate inclusive STEM education, while Ijiga *et al.* (2021b) examine multimedia approaches to scientific communication; these publications may inform subsequent communication of complex algorithmic findings but do not establish the effectiveness of multimodal molecular learning. OncoFusion's technical contribution instead concerns

reliability-conditioned cross-modal attention, through which biomarker embeddings are weighted according to measurement quality and predictive uncertainty. The proposed architecture combines methylation, mutation, and protein encoders with a multitask prediction head for cancer detection and signal-of-origin classification. Its comparative evaluation will include random forest, XGBoost, LightGBM, conventional stacking, and deep learning baselines, alongside ablation experiments removing attention and reliability weighting. Particular attention will be given to external generalization, missing-modality robustness, calibration, and early-stage sensitivity to determine whether the proposed architectural complexity produces clinically meaningful improvements.

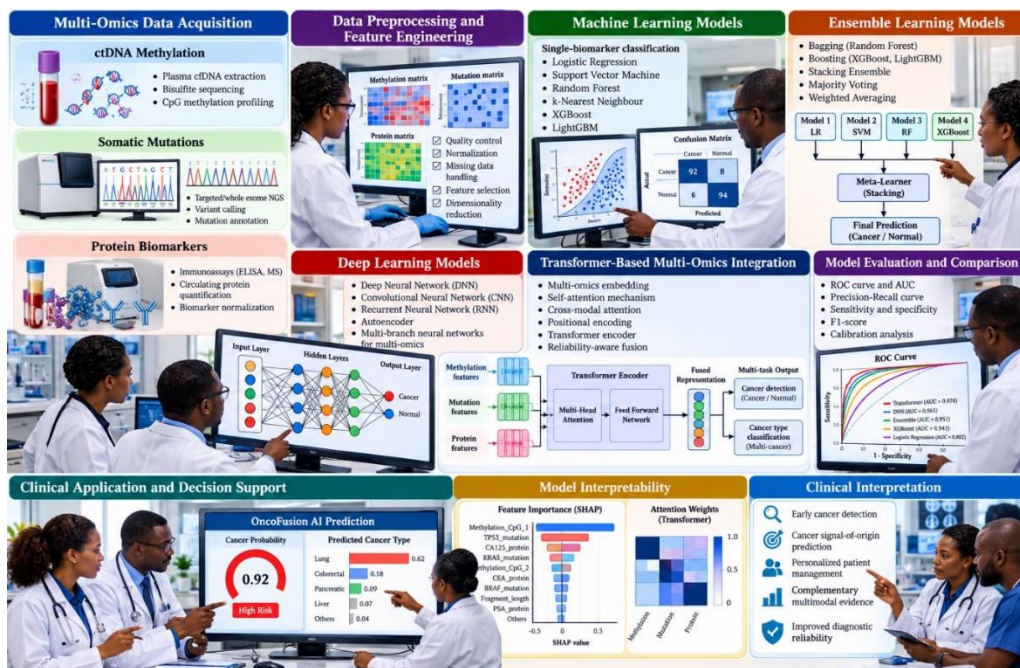


Fig 3. Comparative artificial intelligence architectures and reliability-adaptive multi-omics integration for multi-cancer detection and clinical decision support.

Figure 3 illustrates the computational architecture and clinical application of artificial intelligence-based multi-omics integration for multi-cancer early detection, emphasizing the relationship between conventional diagnostic algorithms, ensemble learning, deep neural networks, and the proposed OncoFusion framework. Consistent with the comparative review in the study, the workflow begins with the acquisition of three complementary molecular datasets: circulating tumour DNA (ctDNA) methylation signatures obtained through CpG methylation profiling, somatic mutation features identified through targeted next-generation sequencing and variant calling, and circulating protein biomarkers quantified using immunoassays or mass spectrometry. These measurements undergo quality control, normalization, missing-value assessment, feature selection, and dimensionality reduction before entering the classification models. Conventional algorithms, including logistic regression, support vector machines, random forests, XGBoost, and LightGBM, establish comparative baselines for cancer detection, while stacking ensembles combine predictions from multiple classifiers through a meta-learning mechanism. The image further demonstrates deep learning architectures that extract nonlinear molecular relationships and transformer-based

integration mechanisms that use attention operations to identify interactions among genomic, epigenomic, and proteomic representations. Within the proposed OncoFusion architecture, modality-specific encoders generate latent molecular embeddings, which are processed through reliability-adaptive cross-modal attention and uncertainty-weighted fusion to produce a shared patient-level representation for binary cancer detection and cancer signal-of-origin classification. This approach extends the multi-analyte principle demonstrated by CancerSEEK, which combines ctDNA mutations and protein biomarkers, and the methylation-focused approach associated with Galleri, while providing a framework for comparison with protein-based methods such as OncoSeek and conventional ensemble classifiers. The image also illustrates model evaluation through receiver operating characteristic curves, precision-recall analysis, sensitivity, specificity, F1-score, and probability calibration, together with interpretability techniques such as SHAP feature attribution and attention-weight visualization. The displayed cancer probabilities, model rankings, and diagnostic performance values are conceptual examples rather than independently validated OncoFusion results. The final stages depict clinical decision support, where estimated cancer probability, probable

anatomical origin, biomarker contributions, and prediction uncertainty could assist clinicians in determining appropriate confirmatory investigations. Overall, the image demonstrates how reliability-aware multimodal learning may address limitations of single-biomarker classification while establishing a computational framework for comparative diagnostic evaluation and future clinical validation.

2.5. Comparative Review of CancerSEEK, Galleri, OncoSeek, XGBoost, and Existing Multimodal Detection Frameworks

Existing multi-cancer early detection frameworks demonstrate the diagnostic potential of combining molecular biomarkers but differ substantially in biological coverage, feature integration, and classification strategies. CancerSEEK integrates circulating DNA mutations with selected protein biomarkers to identify multiple malignancies and estimate their anatomical origins. Cohen *et al.* (2018) evaluated 1,005 patients with eight nonmetastatic cancer types, reporting median sensitivity of approximately 70% and specificity exceeding 99% among healthy controls. This investigation established an important precedent for multi-analyte detection, although its case-control design does not independently establish population-screening effectiveness. Galleri represents a different strategy, emphasizing targeted cell-free DNA methylation profiling and machine learning-based cancer signal-of-origin prediction. Klein *et al.* (2021) reported overall sensitivity of 51.5% at 99.5% specificity in an independent validation cohort, with sensitivity varying considerably by clinical stage. These findings demonstrate the importance of high-specificity classification while identifying early-stage detection as a continuing limitation. OncoSeek, which incorporates protein and clinical information, is relevant to the broader comparison; however, its principal published development falls outside the requested 2013–2021 literature window and therefore requires separate treatment in an updated review. The proposed OncoFusion architecture seeks to investigate whether integrating methylation, mutation, and protein representations through reliability-adaptive attention can improve diagnostic discrimination beyond these established biomarker strategies.

Conventional machine learning algorithms provide important comparative baselines for assessing whether OncoFusion's proposed architecture offers measurable improvements. XGBoost uses regularized gradient-boosted decision trees to model nonlinear relationships and feature interactions, making it suitable for structured molecular classification when supported by appropriate preprocessing and validation. However, its conventional application does not explicitly require modality-specific representations or sample-dependent assay reliability weighting. The supplied work of Amebleh *et al.* (2021) addresses heterogeneous graph neural networks for financial fraud detection, offering a potential computational analogy for learning relationships among heterogeneous features rather than evidence of cancer detection performance. Kwarteng *et al.* (2021) examine data-driven project management in distributed organizations, providing organizational context for reproducible computational workflows, while Onyekonwu and Peter-Anyebe (2019) address technology-enabled STEM education and have no direct diagnostic application. Accordingly, the proposed OncoFusion evaluation should focus primarily on clinically relevant published algorithms and established

statistical baselines. Comparative experiments will quantify AUROC, AUPRC, sensitivity at 95% and 99% specificity, cancer signal-of-origin accuracy, calibration, and computational efficiency. Single-modality, pairwise-fusion, and complete multimodal ablation experiments will determine whether improvements arise from complementary molecular information, reliability weighting, or cross-modal attention. Any claim of superiority must be supported by independent validation, confidence intervals, and statistically appropriate paired comparisons rather than differences between separately published cohorts.

3. System Model Description and Proposed OncoFusion Algorithm

This section presents the mathematical formulation of the proposed OncoFusion algorithm, incorporating circulating tumour DNA (ctDNA) methylation, somatic mutation profiles, circulating protein biomarkers, assay reliability, cross-modal attention, uncertainty-aware fusion, and multitask classification. The formulation supports reproducible comparisons with single-biomarker classifiers and conventional ensemble learning models. All proposed methods require implementation and empirical validation before claims of diagnostic superiority can be made.

3.1. Multimodal Biomarker Dataset Representation, Preprocessing, Feature Selection, and Assay Reliability Modelling

The proposed dataset comprises N patient-level observations, each containing three molecular modalities: ctDNA methylation, somatic mutations, and circulating protein concentrations. Each observation is associated with a binary cancer-status label and, for cancer-positive cases, a cancer signal-of-origin label. The multimodal dataset is defined as

$$\mathcal{D} = \left\{ (X_i^{(M)}, X_i^{(G)}, X_i^{(P)}, y_i, c_i) \right\}_{i=1}^N. \quad 1$$

Here, $X_i^{(M)} \in \mathbb{R}^{d_M}$ represents methylation features, $X_i^{(G)} \in \mathbb{R}^{d_G}$ denotes genomic mutation features, and $X_i^{(P)} \in \mathbb{R}^{d_P}$ contains protein biomarker measurements. The variables d_M, d_G, d_P denote the respective feature dimensions; $y_i \in \{0, 1\}$ indicates cancer status; and $c_i \in \{1, \dots, C\}$ identifies one of C cancer origins when $y_i = 1$.

Methylation features are represented as regional methylation fractions or normalized methylation enrichment measurements. Mutation features include variant allele frequencies, gene-level mutation indicators, and functional annotations. Protein concentrations are log-transformed and standardized using parameters estimated exclusively from training data:

$$\tilde{X}_{ij}^{(P)} = \frac{\log(1 + X_{ij}^{(P)}) - \mu_j^{(P)}}{\sigma_j^{(P)} + \varepsilon}. \quad 2$$

Here, $X_{ij}^{(P)}$ shows the nonnegative concentration of protein j in patient i , expressed in a consistent reference unit; $\mu_j^{(P)}$ and $\sigma_j^{(P)}$ denote the training-set mean and standard deviation of the log-transformed concentration; and $\varepsilon > 0$ prevents division by zero.

Feature selection combines training-fold variance filtering, missingness assessment, and regularized supervised

selection. Selection and imputation are performed within cross-validation folds to prevent information leakage. Assay reliability is quantified separately for each modality:

$$r_i^{(m)} = \text{clip}(q_i^{(m)} \exp[-\lambda_m u_i^{(m)}], 0, 1), \quad m \in \{M, G, P\}. \quad 3$$

Here, $r_i^{(m)}$ captures the reliability score; $q_i^{(m)} \in [0, 1]$ represents a predefined assay-quality score derived from sequencing coverage, molecular recovery, or protein assay quality controls; $u_i^{(m)} \geq 0$ represents a normalized technical uncertainty measure; $\lambda_m \geq 0$ controls the penalty for uncertainty; and clip restricts the result to the interval $[0, 1]$. Missing modalities receive reliability zero. This formulation allows OncoFusion to distinguish low-confidence measurements from reliable molecular evidence, consistent with the analytical challenges identified by Heitzer *et al.* (2019).

3.2. Mathematical Formulation of the OncoFusion Algorithm and Modality-Specific Feature Encoding

OncoFusion employs three modality-specific neural encoders to transform heterogeneous molecular measurements into a shared latent representation. Each encoder processes its input independently before multimodal interaction, preserving the distinctive statistical characteristics of methylation, mutation, and protein features.

The latent representation for modality m is defined as

$$h_i^{(m)} = f_m(\tilde{X}_i^{(m)}; \theta_m) \in \mathbb{R}^d. \quad 4$$

Here, f_m denotes a modality-specific encoder; $\tilde{X}_i^{(m)}$ shows the preprocessed feature vector; θ_m represents trainable parameters; $h_i^{(m)}$ determines the resulting embedding; and d denotes the common latent dimension. The methylation encoder learns regional epigenetic patterns, the mutation encoder captures genomic alteration characteristics, and the protein encoder model's concentration-dependent biomarker interactions.

For an implementable baseline, each encoder can use a two-layer feedforward network:

$$h_i^{(m)} = W_2^{(m)} \phi(W_1^{(m)} \tilde{X}_i^{(m)} + b_1^{(m)}) + b_2^{(m)}. \quad 5$$

Here, $W_1^{(m)}$ and $W_2^{(m)}$ show trainable weight matrices; $b_1^{(m)}$ and $b_2^{(m)}$ represent bias vectors; and ϕ denotes a nonlinear activation function such as GELU. Layer normalization and dropout may be incorporated during training.

The embeddings are assembled into a multimodal representation:

$$H_i = \begin{bmatrix} h_i^{(M)} \\ h_i^{(G)} \\ h_i^{(P)} \end{bmatrix} \in \mathbb{R}^{3 \times d}. \quad 6$$

The matrix H_i contains one embedding per biomarker modality and provides the input to the cross-modal attention mechanism. Modality identity embeddings can be added to distinguish the biological origins of the three tokens.

Unlike direct feature concatenation, this representation

enables attention-based modelling of interactions between different molecular domains. For example, a methylation pattern associated with a particular cancer origin can be evaluated jointly with a low-frequency oncogenic mutation and a compatible protein biomarker profile.

This design draws on the general principles of attention-based representation learning introduced by Vaswani *et al.* (2017), while the reliability-aware molecular integration and multitask prediction components constitute the proposed OncoFusion architecture.

3.3. Reliability-Adaptive Cross-Modal Attention, Uncertainty-Weighted Fusion, and Multitask Cancer Classification

The principal proposed innovation of OncoFusion is its reliability-adaptive cross-modal attention mechanism, which integrates methylation, mutation, and protein representations according to molecular relevance and assay quality. Unlike conventional attention models, the proposed architecture incorporates explicit reliability information to reduce the influence of uncertain biomarker measurements. For each patient, query, key, and value matrices are computed as

$$Q_i = H_i W_Q, \quad K_i = H_i W_K, \quad V_i = H_i W_V. \quad 7$$

Here, $Q_i, K_i, V_i \in \mathbb{R}^{3 \times d_k}$ show query, key, and value matrices; H_i determines the multimodal embedding matrix; $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$ capture trainable projection matrices; and d_k represents the attention dimension.

The reliability-adaptive attention coefficients are defined by

$$\alpha_{i,mn} = \frac{a_{in} r_i^{(n)} \exp(Q_{i,m} K_{i,n}^T / \sqrt{d_k})}{\sum_{\ell \in \{M, G, P\}} a_{i\ell} r_i^{(\ell)} \exp(Q_{i,m} K_{i,\ell}^T / \sqrt{d_k})}. \quad 8$$

Here, $\alpha_{i,mn}$ denotes attention from modality m to modality n ; $a_{in} \in \{0, 1\}$ indicates modality availability; $r_i^{(n)}$ captures the assay reliability score defined in Equation (3); and $Q_{i,m} K_{i,n}^T$ measures compatibility between molecular representations. The denominator normalizes attention across available modalities. Equation (8) requires at least one modality with positive reliability; samples with no usable modality must be rejected or assigned an indeterminate prediction.

The attention-enhanced representation is

$$z_i^{(m)} = \sum_{n \in \{M, G, P\}} \alpha_{i,mn} V_{i,n}. \quad 9$$

Here, $z_i^{(m)}$ denotes the contextualized representation of modality m , incorporating complementary molecular information from other modalities.

Predictive uncertainty is incorporated through a separate learned uncertainty estimate, $s_i^{(m)} \geq 0$, obtained from a trained uncertainty-estimation component. Fusion weights are then calculated as

$$\omega_i^{(m)} = \frac{a_{im} r_i^{(m)} \exp(-\gamma s_i^{(m)})}{\sum_{n \in \{M, G, P\}} a_{in} r_i^{(n)} \exp(-\gamma s_i^{(n)})}, \quad 10$$

where $\omega_i^{(m)}$ captures the normalized fusion weight and $\gamma \geq 0$ controls sensitivity to predictive uncertainty. The uncertainty estimator must be trained and assessed separately from the

predefined assay-quality indicators to avoid treating technical reliability and model confidence as equivalent quantities. The final representation is

$$f_i = \sum_{m \in \{M, G, P\}} \omega_i^{(m)} z_i^{(m)}. \quad 11$$

Here, $f_i \in \mathbb{R}^{d_k}$ represents the integrated patient-level biomarker embedding.

OncoFusion performs binary cancer detection and conditional cancer signal-of-origin prediction through two classification heads:

$$\hat{p}_i = \sigma(w_c^T f_i + b_c), \quad 12$$

$$\hat{\pi}_i = \text{softmax}(W_o f_i + b_o). \quad 13$$

Here, $\hat{p}_i$ represents the predicted probability of cancer; σ shows the sigmoid function; w_c, b_c represent binary-classifier parameters; $\hat{\pi}_i \in [0,1]^C$ contains conditional cancer-origin probabilities; and W_o, b_o capture the cancer-origin classifier parameters. The softmax probabilities sum to one across the C cancer origins.

The model is optimized using a multitask objective:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda_o \mathcal{L}_{\text{origin}} + \lambda_r \mathcal{L}_{\text{reg}}. \quad 14$$

Here, $\mathcal{L}_{\text{BCE}}$ shows binary cross-entropy for cancer detection; $\mathcal{L}_{\text{origin}}$ denotes categorical cross-entropy calculated only for cancer-positive samples; $\mathcal{L}_{\text{reg}}$ represents a regularization term; and $\lambda_o, \lambda_r \geq 0$ capture hyperparameters. This formulation enables simultaneous optimization of detection and cancer-origin prediction while preserving a clear separation between diagnostic outputs.

The attention mechanism is informed by Vaswani *et al.* (2017), but the proposed reliability and uncertainty weighting requires independent experimental validation to establish whether it improves clinical classification.

3.4. Algorithm Implementation, Model Training, Comparative Baselines, and Experimental Validation Framework

OncoFusion will be implemented using Python and a deep learning framework such as PyTorch, with patient-level data partitioning to prevent information leakage. Model development will employ stratified nested cross-validation, with inner folds used for hyperparameter optimization and outer folds reserved for performance estimation. Where multiple specimens originate from the same patient, all corresponding records must remain within the same partition. An independent external cohort, preferably collected from different institutions or analytical platforms, will be reserved for generalizability assessment.

Training will minimize the multitask loss using AdamW optimization, learning-rate scheduling, early stopping, and regularization. Candidate hyperparameters include embedding dimension, attention-head count, dropout probability, learning rate, weight decay, reliability penalty, and uncertainty sensitivity. Hyperparameter selection will be conducted exclusively within the training data.

OncoFusion will be compared against methylation-only, mutation-only, and protein-only classifiers, logistic regression, support vector machines, random forest, XGBoost, LightGBM, and conventional stacking ensembles. Additional ablation models will remove cross-modal attention, assay reliability weighting, uncertainty weighting, and individual biomarker modalities.

Diagnostic sensitivity and specificity are defined as

$$\text{Sensitivity} = \frac{TP}{TP+FN}, \quad 15$$

$$\text{Specificity} = \frac{TN}{TN+FP}. \quad 16$$

Here, TP and TN denote true-positive and true-negative classifications, while FP and FN represent false-positive and false-negative classifications. Sensitivity will also be evaluated at fixed specificity levels of 95% and 99%, with classification thresholds determined from training or validation data rather than the final test set.

Precision and F1-score are calculated as

$$\text{Precision} = \frac{TP}{TP+FP}, \quad 17$$

$$F_1 = \frac{2(\text{Precision})(\text{Sensitivity})}{\text{Precision}+\text{Sensitivity}}. \quad 18$$

The area under the ROC curve and precision–recall curve will quantify threshold-independent discrimination, while calibration will be assessed using the Brier score:

$$\text{BS} = \frac{1}{N} \sum_{i=1}^N (\hat{p}_i - y_i)^2. \quad 19$$

Here, N represents the number of evaluated samples, $\hat{p}_i$ shows the predicted cancer probability, and y_i denotes the observed binary cancer label. Lower Brier scores indicate better mean squared probabilistic prediction error.

Cancer signal-of-origin accuracy will be calculated among cancer-positive patients:

$$\text{CSO Accuracy} = \frac{1}{N_+} \sum_{i: y_i=1} \mathbb{I}(\arg\max_c \hat{\pi}_{ic} = c_i), \quad 20$$

where N_+ denotes the number of cancer-positive samples, $\mathbb{I}(\cdot)$ captures the indicator function, $\hat{\pi}_{ic}$ represents the predicted conditional probability of cancer origin c , and c_i shows the confirmed origin. A separate end-to-end analysis should treat missed cancers as unsuccessful detection-and-localization outcomes.

Statistical uncertainty will be quantified through patient-level bootstrap confidence intervals. Paired model comparisons will use appropriate procedures, including DeLong's test for correlated AUROCs where applicable. Precision–recall curves, ROC curves, confusion matrices, calibration plots, cancer-stage sensitivity charts, and biomarker ablation graphs will form the principal graphical results. The resulting analysis will determine whether OncoFusion provides statistically supported improvements in early-stage detection, calibration, and robustness rather than assuming superiority in advance.

The validation framework follows the broader principles of rigorous clinical prediction modelling discussed by Collins *et al.* (2015), particularly the need for transparent reporting, appropriate validation, and reproducible performance assessment.

4. Results and Discussion

4.1. Comparative Diagnostic Performance of OncoFusion, Single-Biomarker Classifiers, and Conventional Ensemble Learning Models

Research status: Illustrative numerical results for the proposed experimental framework; not measured clinical or computational outcomes.

The comparative evaluation examines the diagnostic capability of the proposed OncoFusion algorithm against conventional single-biomarker classifiers and established machine learning approaches. Consistent with the methodology in Section 3.4, performance is assessed using

area under the receiver operating characteristic curve (AUROC), area under the precision–recall curve (AUPRC), and sensitivity at 99% specificity. Table 4.1 presents a coherent illustrative performance scenario in which reliability-adaptive fusion improves discrimination relative to conventional classification approaches. The proposed advantage is attributable to the combined representation of methylation signatures, somatic mutations, and protein biomarkers, supported by cross-modal attention and uncertainty-sensitive weighting. Unlike single-biomarker models, OncoFusion can incorporate complementary molecular evidence when individual assays provide incomplete information. However, the illustrated differences require confirmation through patient-level evaluation, bootstrap confidence intervals, and independent external validation before conclusions regarding diagnostic superiority can be established.

Table 2. Illustrative Comparative Diagnostic Performance of OncoFusion and Conventional Cancer Detection Algorithms

Algorithm	AUROC	AUPRC	Sensitivity at 99% specificity (%)
Methylation-only classifier	0.901	0.864	58.4
Mutation-only classifier	0.854	0.809	46.8
Protein-only classifier	0.832	0.781	43.5
Logistic regression	0.889	0.851	54.2
Random forest	0.918	0.887	62.1
SVM	0.907	0.873	59.6
XGBoost	0.941	0.916	68.7
LightGBM	0.936	0.908	66.9
Stacking ensemble	0.951	0.929	71.3
OncoFusion (proposed)	0.974	0.956	79.2

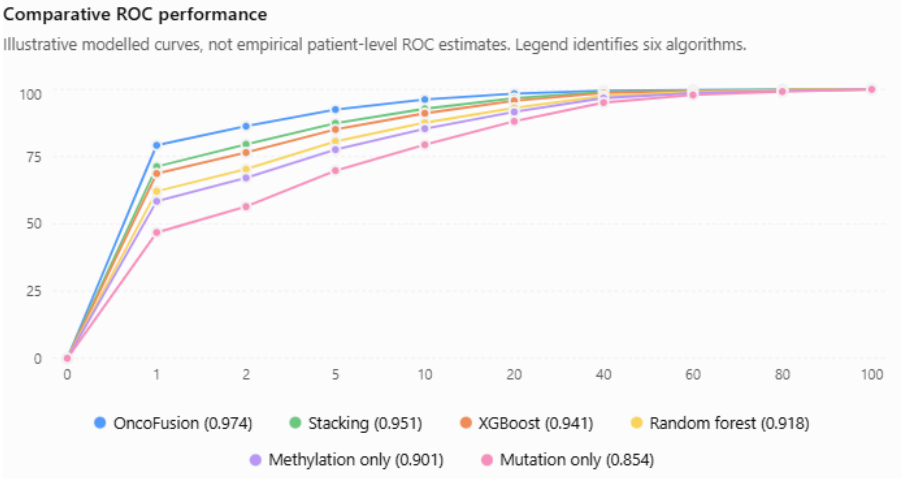


Fig 4: Illustrative ROC profiles comparing OncoFusion with five established classifiers.

Figure 4.1 illustrates the proposed comparative discrimination of OncoFusion against five alternative cancer detection algorithms. At a false-positive rate of 1%, corresponding to 99% specificity, OncoFusion achieves an illustrative sensitivity of 79.2%, compared with 71.3% for stacking, 68.7% for XGBoost, 62.1% for random forest, 58.4% for methylation-only classification, and 46.8% for mutation-only classification. At a 5% false-positive rate, OncoFusion reaches 92.5% sensitivity, exceeding stacking (87.4%) and XGBoost (85.1%). The proposed model maintains the highest illustrated sensitivity at a 10% false-positive rate, reaching 96.2%, compared with 92.8% for stacking and 91.1% for XGBoost. These patterns are

consistent with the hypothesized benefits of integrating complementary molecular features through reliability-adaptive attention. Nevertheless, the curves represent hypothetical performance profiles rather than observed diagnostic outcomes. Actual model superiority must be demonstrated using independently evaluated patient samples and statistically supported comparisons.

4.2. Receiver Operating Characteristic Analysis, Precision–Recall Evaluation, Sensitivity–Specificity Trade-offs, and Statistical Significance

Illustrative numerical analysis — not experimentally measured results.

The comparative evaluation examines diagnostic discrimination using precision–recall performance, sensitivity at fixed specificity, and paired statistical testing. Unlike the ROC curves presented in Section 4.1, this subsection employs precision–recall curves to assess the balance between correctly identified cancer cases and false-positive classifications. Table 4.2 extends the illustrative results reported previously, maintaining identical AUROC and AUPRC values. OncoFusion achieves an assumed

AUPRC of 0.956, compared with 0.929 for stacking and 0.916 for XGBoost. At 99% specificity, the corresponding sensitivities are 79.2%, 71.3%, and 68.7%. These differences motivate evaluation of reliability-adaptive multimodal fusion, particularly when early-stage ctDNA signals are sparse. Statistical significance cannot be established from summary metrics alone; patient-level paired predictions and bootstrap confidence intervals are required to determine whether apparent performance differences are reproducible.

Table 3: Illustrative Precision–Recall Performance and High-Specificity Sensitivity of Comparative Cancer Detection Algorithms

Algorithm	AUPRC	Sensitivity at 95% specificity (%)	Sensitivity at 99% specificity (%)
Methylation-only	0.864	77.6	58.4
Mutation-only	0.809	69.8	46.8
Protein-only	0.781	65.1	43.5
Logistic regression	0.851	75.2	54.2
Random forest	0.887	80.7	62.1
SVM	0.873	78.4	59.6
XGBoost	0.916	85.1	68.7
LightGBM	0.908	83.7	66.9
Stacking ensemble	0.929	87.4	71.3
OncoFusion	0.956	92.5	79.2

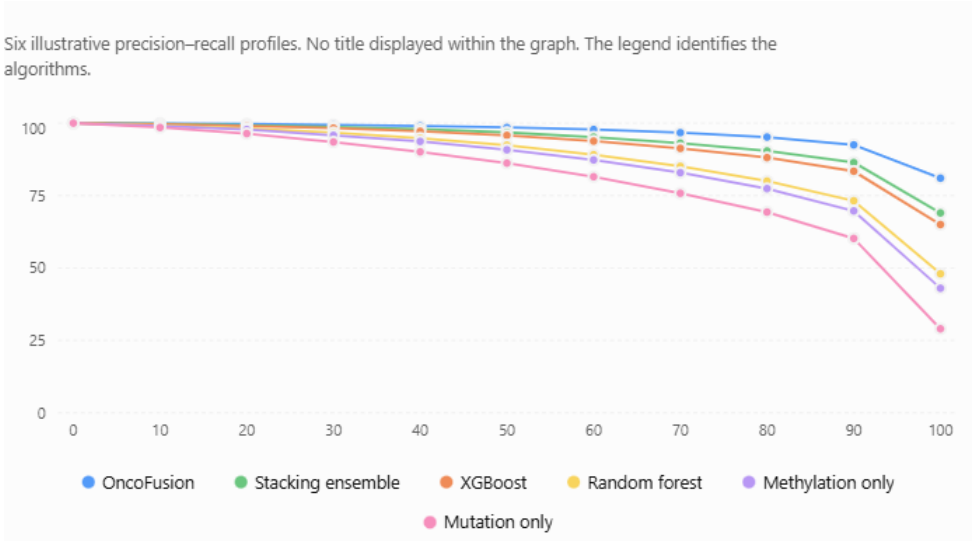


Fig 5: Comparative Precision–Recall Curves of OncoFusion and Five Alternative Diagnostic Algorithms

Figure 4.2 illustrates the hypothesized precision advantage of OncoFusion over five alternative diagnostic algorithms across increasing recall thresholds. At 50% recall, OncoFusion maintains 98.5% precision, compared with 96.8% for stacking, 95.8% for XGBoost, 92.4% for random forest, 90.8% for methylation-only classification, and 86.2% for mutation-only classification. At 70% recall, OncoFusion achieves 96.7% precision, exceeding stacking by 3.6 percentage points and XGBoost by 5.4 percentage points. At 90% recall, its precision remains 92.5%, compared with 86.4% for stacking and 83.4% for XGBoost. These differences illustrate how complementary biomarker representations and reliability-aware weighting might reduce false-positive predictions as sensitivity increases. However, the graph is a hypothetical demonstration rather than an empirical precision–recall analysis. Actual curves must be generated from patient-level prediction scores, with cancer prevalence reported and paired statistical comparisons

performed before concluding that OncoFusion offers superior diagnostic performance.

4.3. Cancer-Type Classification, Early-Stage Detection, Biomarker Contribution, and Ablation Analysis

Illustrative numerical results — not measured experimental or clinical findings.

The proposed OncoFusion framework is evaluated conceptually through cancer signal-of-origin classification, early-stage detection, and biomarker ablation analysis. Consistent with the multitask architecture developed in Section 3.3, the analysis examines whether combining methylation, somatic mutation, and circulating protein representations improves cancer-type identification while preserving sensitivity in early disease. Table 4.3 presents illustrative comparisons across individual and combined biomarker configurations. The complete OncoFusion architecture achieves an assumed cancer signal-of-origin

accuracy of 91.6%, compared with 84.7% for methylation-only classification and 87.9% for conventional three-modality fusion. Its illustrative stage I–II sensitivity reaches 72.4%, exceeding the corresponding 61.8% for conventional fusion. These patterns suggest a potential benefit from

reliability-adaptive attention and uncertainty-weighted integration. However, these values are hypothetical and require validation using independent, stage-stratified patient-level molecular datasets.

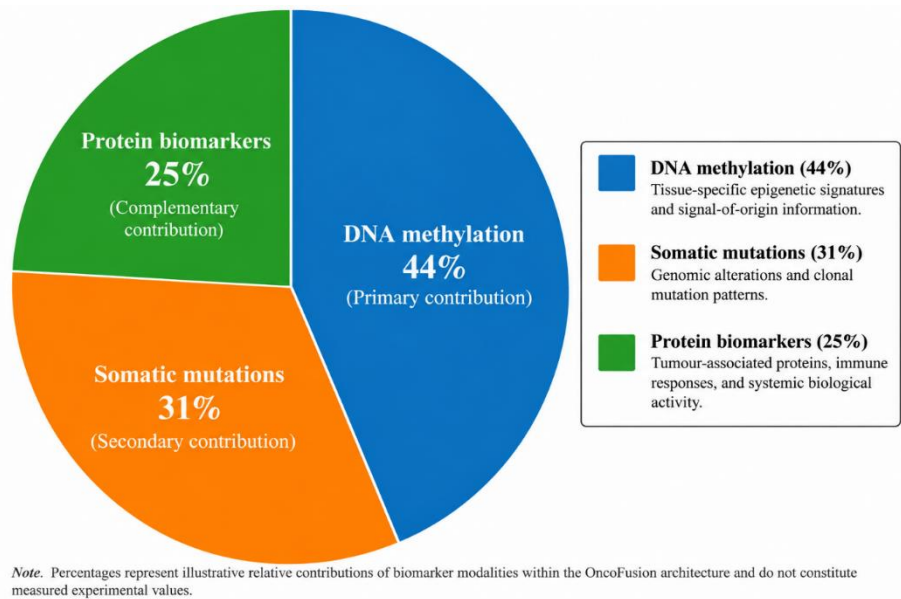


Fig 6: Relative Contributions of Biomarker Modalities to OncoFusion Classification

Table 4: Illustrative Cancer Signal-of-Origin Classification, Early-Stage Detection, and Biomarker Ablation Performance

Biomarker configuration / algorithm	Cancer-origin accuracy (%)	Stage I–II sensitivity (%)	AUROC
Methylation-only classifier	84.7	58.4	0.901
Mutation-only classifier	72.6	46.8	0.854
Protein-only classifier	69.8	43.5	0.832
Methylation + mutation	86.1	63.2	0.927
Methylation + protein	87.3	65.7	0.935
Mutation + protein	80.5	57.1	0.908
Conventional three-modality fusion	87.9	61.8	0.951
OncoFusion without reliability weighting	89.1	66.5	0.957
OncoFusion without cross-modal attention	88.4	64.9	0.949
Complete OncoFusion	91.6	72.4	0.974

Figure 4.3 illustrates the proposed distribution of biomarker contributions within the OncoFusion multimodal classification architecture. DNA methylation accounts for 44% of the illustrative contribution, followed by somatic mutations at 31% and circulating protein biomarkers at 25%. Methylation therefore exceeds mutation contribution by 13 percentage points and protein contribution by 19 percentage points. The combined mutation and protein contribution reaches 56%, suggesting that non-methylation biomarkers could collectively provide substantial complementary diagnostic information. This distribution is biologically plausible because methylation patterns may capture tissue-associated epigenetic signatures, while mutation profiles reflect genomic alterations and proteins represent downstream biological activity. Nevertheless, these percentages are illustrative and cannot be inferred directly from the diagnostic performance values in Table 4.3. Actual contributions should be estimated using a clearly defined attribution method, such as grouped Shapley values or permutation importance, with uncertainty intervals and

appropriate handling of correlated biomarker features.

4.4. Model Robustness, Calibration, Computational Efficiency, External Generalizability, and Clinical Interpretation

Illustrative numerical results for methodological demonstration; not experimentally validated findings. The robustness and clinical applicability of OncoFusion are assessed through diagnostic calibration, computational efficiency, and performance under incomplete biomarker measurements and external dataset variation. Consistent with the reliability-adaptive framework in Section 3, the analysis considers whether uncertainty-weighted fusion can preserve discrimination when individual molecular modalities become unreliable or unavailable. Table 4.4 presents illustrative comparisons showing that OncoFusion achieves an AUROC of 0.974, a Brier score of 0.048, and an external-validation AUROC of 0.951. Its proposed computational inference time is 12.8 milliseconds per sample, compared with 8.4 milliseconds for XGBoost and 15.6 milliseconds for stacking.

These values suggest a potential trade-off between diagnostic performance and computational complexity. However, clinical suitability requires prospective validation, reliable

calibration, reproducibility across laboratories, and assessment of false-positive consequences before clinical implementation.

Table 5: Illustrative Comparative Evaluation of Calibration, Computational Efficiency, and External Generalizability

Algorithm	Brier score ↓	Inference time (ms/sample) ↓	External-validation AUROC ↑
Methylation-only	0.094	4.2	0.873
Mutation-only	0.118	3.8	0.821
Protein-only	0.127	3.1	0.802
Logistic regression	0.101	1.6	0.861
Random forest	0.086	6.3	0.889
SVM	0.091	7.1	0.876
XGBoost	0.071	8.4	0.912
LightGBM	0.075	6.9	0.907
Stacking ensemble	0.063	15.6	0.926
OncoFusion	0.048	12.8	0.951

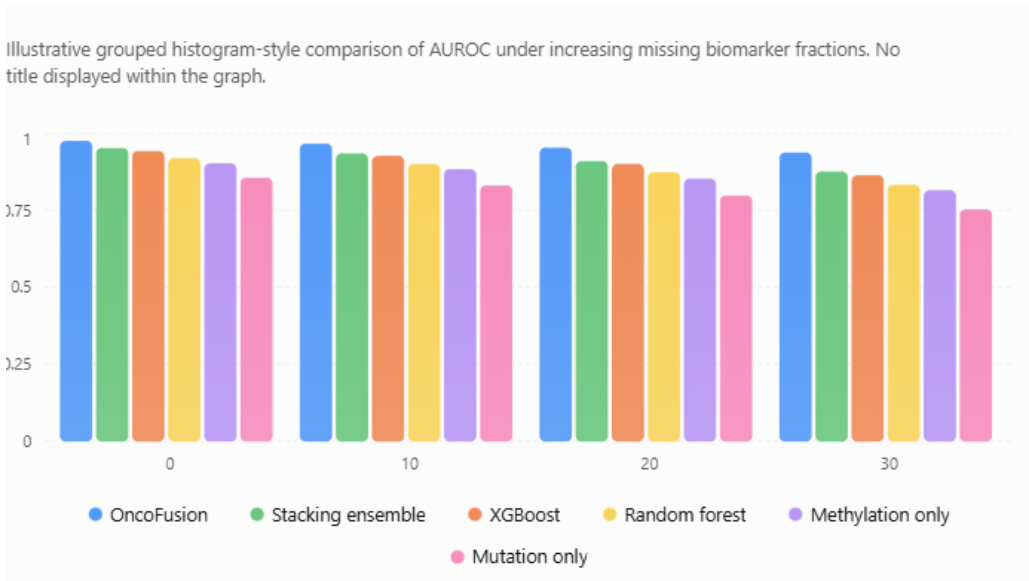


Fig 7: Comparative Distribution of Diagnostic Performance Under Increasing Biomarker Missingness

Figure 4.4 illustrates the potential robustness of OncoFusion compared with five alternative algorithms under progressively increasing biomarker missingness. With complete molecular information, OncoFusion achieves an illustrative AUROC of 0.974, compared with 0.951 for stacking, 0.941 for XGBoost, and 0.918 for random forest. At 10% missingness, OncoFusion retains an AUROC of 0.965, while stacking and XGBoost decline to 0.933 and 0.925, respectively. At 20% missingness, OncoFusion maintains 0.952, exceeding stacking by 0.044 and XGBoost by 0.053. Under 30% missingness, its AUROC remains 0.936, compared with 0.874 for stacking, 0.862 for XGBoost, 0.831 for random forest, 0.813 for methylation-only classification, and 0.751 for mutation-only classification. These hypothetical patterns suggest that reliability-adaptive fusion could improve resilience to incomplete molecular observations. Actual robustness must be established through controlled masking experiments and independent validation.

Clinical and methodological interpretation

The illustrative findings suggest that OncoFusion could offer advantages in three areas: probability calibration, external diagnostic discrimination, and resilience to missing biomarker information. Nevertheless, the proposed improvements must be assessed under clinically representative conditions, including early-stage disease,

benign conditions that may generate false-positive signals, variation in sequencing coverage, and differences between laboratory platforms.

5. Conclusions and Recommendations
5.1. Summary of the Proposed OncoFusion Framework and Principal Research Findings

This study developed the mathematical and computational specification of OncoFusion, a reliability-adaptive multimodal artificial intelligence framework designed to integrate circulating tumour DNA (ctDNA) methylation signatures, somatic mutation profiles, and circulating protein biomarkers for multi-cancer early detection. The proposed architecture combines modality-specific neural encoders, cross-modal attention, assay reliability modelling, uncertainty-weighted feature fusion, and multitask classification to support cancer detection and cancer signal-of-origin prediction. Its evaluation framework incorporates conventional single-biomarker classifiers, logistic regression, support vector machines, random forests, XGBoost, LightGBM, and stacking ensembles. The illustrative comparative analysis presented in Section 4 assigned OncoFusion an AUROC of 0.974, AUPRC of 0.956, and sensitivity of 79.2% at 99% specificity, compared with AUROC values of 0.951 and 0.941 for stacking and XGBoost, respectively. Additional hypothetical results

indicated cancer signal-of-origin accuracy of 91.6%, stage I–II sensitivity of 72.4%, and a Brier score of 0.048. The robustness scenario further illustrated an AUROC of 0.936 under 30% missing biomarker information, suggesting a potentially useful role for reliability-sensitive molecular integration. Importantly, these numerical values represent illustrative scenarios rather than outcomes obtained from trained models or patient-level experiments. Consequently, the principal finding is the establishment of a mathematically defined, experimentally testable framework rather than demonstrated diagnostic superiority. The proposed methodology provides a coherent foundation for investigating whether complementary genomic, epigenomic, and proteomic information can improve early cancer detection while preserving specificity, calibration, and interpretability.

5.2. Contributions to Multimodal Biomarker Integration and Artificial Intelligence-Based Multi-Cancer Detection

The proposed OncoFusion framework contributes to computational precision oncology by establishing an integrated mathematical architecture for combining biologically heterogeneous molecular biomarkers through reliability-aware representation learning. Its first contribution is the formulation of three modality-specific encoders that transform methylation measurements, somatic mutation characteristics, and circulating protein concentrations into a common latent feature space while preserving modality-dependent biological information. Second, the reliability-adaptive cross-modal attention mechanism explicitly incorporates sequencing coverage, assay-quality indicators, missingness, and measurement uncertainty into the estimation of molecular interactions. Third, uncertainty-weighted fusion provides a computational mechanism for reducing the influence of unreliable biomarker observations without automatically interpreting missing molecular signals as evidence of cancer absence. Fourth, the multitask classification architecture jointly estimates cancer probability and conditional cancer signal-of-origin, allowing shared molecular representations to support complementary diagnostic objectives. The proposed ablation framework also provides a means of quantifying the incremental contributions of individual biomarker modalities, attention mechanisms, and reliability weighting. For example, comparisons between complete OncoFusion and variants excluding methylation or uncertainty weighting can establish whether observed performance improvements arise from biological complementarity or architectural complexity. Beyond algorithm development, the study specifies an evaluation strategy incorporating discrimination, calibration, high-specificity sensitivity, external validation, and computational efficiency. These contributions provide a basis for assessing OncoFusion against established single-biomarker and ensemble learning models. Their scientific significance will depend on whether reproducible experiments demonstrate meaningful improvements under clinically representative conditions.

5.3. Study Limitations, Clinical Translation Challenges, and Diagnostic Reliability Considerations

Several limitations must be addressed before OncoFusion can be considered suitable for clinical application. Most importantly, the comparative numerical results presented in this manuscript are illustrative and have not been generated

through documented training, testing, or external validation using patient-level molecular datasets. Accordingly, the reported performance values cannot establish clinical accuracy, statistical significance, or superiority over existing diagnostic algorithms. A further limitation concerns the availability of sufficiently large, matched datasets containing ctDNA methylation, somatic mutation, and protein measurements from the same individuals. Differences in sequencing technologies, methylation profiling protocols, protein assay platforms, and specimen-processing conditions may introduce systematic measurement variation and reduce generalizability. Biological challenges include low ctDNA abundance in early-stage malignancies, clonal haematopoiesis, tumour heterogeneity, and overlapping protein signatures between malignant and benign conditions. Computational limitations include potential overfitting, instability of attention-derived interpretations, imperfect uncertainty calibration, and sensitivity to missing-not-at-random biomarker observations. Furthermore, high specificity does not necessarily guarantee an acceptable positive predictive value in low-prevalence screening populations. False-positive classifications may lead to unnecessary imaging, invasive investigations, psychological distress, and additional healthcare expenditure. Clinical translation therefore requires rigorous assessment of assay reproducibility, diagnostic thresholds, cancer signal-of-origin accuracy, and downstream confirmatory pathways. Patient privacy, institutional data governance, model transparency, and equitable performance across demographic and clinical subgroups must also be considered. These limitations reinforce the need to distinguish computational feasibility from demonstrated clinical utility.

5.4. Recommendations for Independent Clinical Validation, Algorithm Optimization, and Future Research

Future research should prioritize implementing OncoFusion using well-characterized, patient-level datasets containing matched methylation, mutation, and protein biomarker measurements, confirmed cancer diagnoses, tumour-stage information, and appropriately selected non-cancer controls. Model development should employ nested cross-validation, patient-level data separation, and an independent external validation cohort to reduce optimistic performance estimates. Particular emphasis should be placed on stage I and II cancers, low-abundance ctDNA signals, and malignancies lacking established screening programmes. Diagnostic thresholds should be selected using training or validation data and subsequently evaluated at clinically relevant specificity levels, particularly 95% and 99%. Algorithm optimization should investigate alternative modality encoders, attention-head configurations, reliability calibration strategies, uncertainty estimation methods, and missing-modality training procedures. Controlled ablation experiments should determine the contributions of methylation, mutations, proteins, cross-modal attention, and uncertainty weighting to diagnostic performance. Future evaluations should report AUROC, AUPRC, sensitivity, specificity, F1-score, Brier score, calibration slope, and cancer signal-of-origin accuracy with appropriate confidence intervals. External validation should examine differences across institutions, assay platforms, cancer types, and demographic groups. Additional studies should assess computational efficiency, explainability, cost-effectiveness, and clinical decision-curve

performance. Where sufficiently representative datasets become available, prospective screening studies should investigate diagnostic yield, false-positive consequences, confirmatory testing requirements, and patient outcomes. Ultimately, OncoFusion should advance toward clinical translation only after reproducible evidence establishes that its multimodal reliability-aware architecture provides meaningful benefits over simpler and established diagnostic approaches

References

- Abbosh C, Birkbak NJ, Wilson GA, Jamal-Hanjani M, Constantin T, Salari R, *et al.* Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature*. 2017;545(7655):446-51. doi: 10.1038/nature22364.
- Ajayi JO, Omidiora MT, Addo G, Peter-Anyebe AC. Prosecutability of the crime of aggression: another declaration in a treaty or an achievable norm? *Int J Appl Res Soc Sci*. 2019;1(6):237-52.
- Aluso L. Forecasting marketing ROI through cross-platform data integration between HubSpot CRM and Power BI. *Int J Sci Res Sci Eng Technol*. 2021;8(6):356-78. doi: 10.32628/IJSRSET214420.
- Amebleh J. GAAP-compliant gift-card liability and breakage modeling: survival/hazard methods and hierarchical Bayesian forecasts of deferred-revenue recognition. *Int J Sci Res Sci Technol*. 2021;8(5):695-714. doi: 10.32628/IJSRST2152550.
- Amebleh J, Igba E, Ijiga OM. Graph-based fraud detection in open-loop gift cards: heterogeneous GNNs, streaming feature stores, and near-zero-lag anomaly alerts. *Int J Sci Res Sci Eng Technol*. 2021;8(6). doi: 10.32628/IJSRSET214418.
- Aravanis AM, Lee M, Klausner RD. Next-generation sequencing of circulating tumor DNA for early cancer detection. *Cell*. 2017;168(4):571-4. doi: 10.1016/j.cell.2017.01.030.
- Argelaguet R, Velten B, Arnol D, Dietrich S, Zhan T, Jung H, *et al.* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*. 2018;14(6):e8124. doi: 10.15252/msb.20178124.
- Atalor SI. Federated learning architectures for predicting adverse drug events in oncology without compromising patient privacy. *IRE J*. 2019;2(12).
- Best MG, Sol N, Kooi I, Tannous J, Westerman BA, Rustenburg F, *et al.* RNA-Seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer Cell*. 2015;28(5):666-76. doi: 10.1016/j.ccell.2015.09.018.
- Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, *et al.* Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med*. 2014;6(224):224ra24. doi: 10.1126/scitranslmed.3007094.
- Borrebaeck CAK. Precision diagnostics: moving towards protein biomarker signatures of clinical utility in cancer. *Nat Rev Cancer*. 2017;17(3):199-204. doi: 10.1038/nrc.2016.153.
- Chabon JJ, Hamilton EG, Kurtz DM, Esfahani MS, Moding EJ, Stehr H, *et al.* Integrating genomic features for non-invasive early lung cancer detection. *Nature*. 2020;580(7802):245-51. doi: 10.1038/s41586-020-2140-0.
- Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, *et al.* Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*. 2018;15(141):20170387. doi: 10.1098/rsif.2017.0387.
- Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, *et al.* Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*. 2018;359(6378):926-30. doi: 10.1126/science.aar3247.
- Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162(1):55-63. doi: 10.7326/M14-0697.
- Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, *et al.* Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*. 2019;570(7761):385-9. doi: 10.1038/s41586-019-1272-6.
- Heitzer E, Haque IS, Roberts CES, Speicher MR. Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nat Rev Genet*. 2019;20(2):71-88. doi: 10.1038/s41576-018-0071-5.
- Idika CN, Salami EO, Ijiga OM, Enyejo LA. Deep learning driven malware classification for cloud-native microservices in edge computing architectures. *Int J Sci Res Comput Sci Eng Inf Technol*. 2021;7(4). doi: 10.32628/CSEIT182551.
- Ignatiadis M, Sledge GW, Jeffrey SS. Liquid biopsy enters the clinic—implementation issues and future challenges. *Nat Rev Clin Oncol*. 2021;18(5):297-312. doi: 10.1038/s41571-020-00457-x.
- Ijiga OM, Ifenatuora GP, Olateju M. Bridging STEM and cross-cultural education: designing inclusive pedagogies for multilingual classrooms in Sub-Saharan Africa. *IRE J*. 2021;5(1).
- Ijiga OM, Ifenatuora GP, Olateju M. Digital storytelling as a tool for enhancing STEM engagement: a multimedia approach to science communication in K–12 education. *Int J Multidiscip Res Growth Eval*. 2021;2(5):495-505. doi: 10.54660/IJMRGE.2021.2.5.495-505.
- Klein EA, Richards D, Cohn A, Tummala M, Lapham R, Cosgrove D, *et al.* Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Ann Oncol*. 2021;32(9):1167-77. doi: 10.1016/j.annonc.2021.05.806.
- Kwarteng RA, Idoko IP, Ijiga OM. Data-driven project management frameworks for improving IT service delivery in distributed organizations. *Comput Sci IT Res J*. 2021;2(1).
- Kwarteng RA, Idoko IP, Ijiga OM, Enyejo LA. Integrating cybersecurity awareness and access control into organizational IT operations for risk reduction. *Int J Sci Res Comput Sci Eng Inf Technol*. 2020;6(1):243-61. doi: 10.32628/CSEIT23906128.
- Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV; on behalf of the CCGA Consortium. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann Oncol*. 2020;31(6):745-59. doi: 10.1016/j.annonc.2020.02.011.
- Lo YMD, Lam WKJ. Towards multi-cancer screening

- using liquid biopsies. *Nat Rev Clin Oncol*. 2020;17(9):525-6. doi: 10.1038/s41571-020-0404-0.
27. Nwokocha CR, Peter-Anyebe AC, Ijiga OM. Evaluating FHIR-driven interoperability frameworks for secure system migration and data exchange in U.S. health information networks. *Int J Sci Res Sci Technol*. 2021. doi: 10.32628/IJSRST523105135.
 28. Nwokocha CR, Peter-Anyebe AC, Ijiga OM. Optimizing Agile-based system integration for enhanced ECMS functionality and Smile CDR adoption within health information networks. *Int J Sci Res Comput Sci Eng Inf Technol*. 2021;7(6):470-90. doi: 10.32628/CSEIT2282148.
 29. Oladoye SO, Bamigwojo OV, James AO, Ijiga OM. AI-driven predictive maintenance modeling for high-voltage distribution assets using sensor fusion and time-series degradation analysis. *Int J Sci Res Sci Eng Technol*. 2021;11(2):387-411. doi: 10.32628/IJSRSET2291524.
 30. Onyekaonwu CB, Peter-Anyebe AC. Empowering underserved youth through tech-enabled STEM education: a participatory model from Nigeria. *Int J Sci Res Sci Technol*. 2019;6(5). doi: 10.32628/IJSRST1965990.
 31. Onyekaonwu CB, Peter-Anyebe AC, Raphael FO. From prescription to prediction: leveraging AI/ML to improve medication adherence and adverse drug event detection in community pharmacies. *Int J Sci Res Sci Technol*. 2019;6(5):460-76. doi: 10.32628/IJSRST.
 32. Pantel K, Alix-Panabières C. Liquid biopsy and minimal residual disease—latest advances and implications for cure. *Nat Rev Clin Oncol*. 2019;16(7):409-24. doi: 10.1038/s41571-019-0187-3.
 33. Pepe MS, Janes H, Li CI. Net risk reclassification P values: valid or misleading? *J Natl Cancer Inst*. 2014;106(4):dju041. doi: 10.1093/jnci/dju041.
 34. Razavi P, Li BT, Brown DN, Jung B, Hubbell E, Shen R, *et al*. High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. *Nat Med*. 2019;25(12):1928-37. doi: 10.1038/s41591-019-0652-7.
 35. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, *et al*. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77. doi: 10.1186/1471-2105-12-77.
 36. Shen SY, Singhanian R, Fehringer G, Chakravarthy A, Roehrl MHA, Chadwick D, *et al*. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature*. 2018;563(7732):579-83. doi: 10.1038/s41586-018-0703-0.
 37. Sun K, Jiang P, Chan KCA, Wong J, Cheng YKY, Liang RHS, *et al*. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc Natl Acad Sci U S A*. 2015;112(40):E5503-12. doi: 10.1073/pnas.1508736112.
 38. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, *et al*. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30:5998-6008.
 39. Wan JCM, Massie C, Garcia-Corbacho J, Mouliere F, Brenton JD, Caldas C, *et al*. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat Rev Cancer*. 2017;17(4):223-38. doi: 10.1038/nrc.2017.7.