



Development of the OncoSignal Algorithm for Detecting Low Abundance Circulating Tumor DNA Signals Compared with Conventional Threshold Based and Machine Learning Detection

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

ISSN (online): 2582-8940

Volume: 03

Issue: 01

January-March 2022

Received: 05-02-2022

Accepted: 03-03-2022

Page No: 87-103

Abstract

Circulating tumor DNA (ctDNA) provides a minimally invasive molecular signal for cancer detection, residual disease assessment, treatment-response monitoring, and recurrence surveillance; however, reliable identification becomes difficult when tumor-derived fragments occur at very low variant allele fractions within a dominant background of normal cell-free DNA. This study develops OncoSignal, a novel computational algorithm for detecting low-abundance ctDNA through integrated background-error modelling, fragment-level signal weighting, molecular-context encoding, adaptive denoising, probabilistic evidence fusion, and confidence-calibrated classification. Unlike conventional fixed-threshold approaches that classify candidate variants primarily according to read depth, allele frequency, or predefined quality cut-offs, OncoSignal constructs a multidimensional evidence score from variant allele frequency, base quality, mapping quality, strand consistency, fragment length, genomic context, local sequencing error, molecular support, and locus-specific noise characteristics. The proposed architecture incorporates an Adaptive Molecular Noise Suppression module, Fragment-Context Attention Network, Bayesian Evidence Fusion layer, and Dynamic Signal Decision Boundary, allowing the detection threshold to adjust to sequencing depth, background error distribution, and sample-specific noise. OncoSignal is designed for comparative benchmarking against conventional variant allele frequency thresholding, statistical binomial detection, logistic regression, support vector machine, random forest, XGBoost, multilayer perceptron, one-dimensional convolutional neural networks, long short-term memory networks, and Transformer-based classifiers. Performance is evaluated across progressively decreasing ctDNA fractions using sensitivity, specificity, precision, F1-score, Matthews correlation coefficient, area under the receiver operating characteristic curve, area under the precision-recall curve, false-positive rate, calibration error, signal-to-noise ratio, and limit of detection. Particular emphasis is placed on the clinically demanding low-abundance interval below 1% variant allele frequency and the ultra-low range below 0.1%. Comparative graphical analysis includes ROC and precision-recall curves, sensitivity-versus-variant-allele-frequency profiles, three-dimensional surfaces relating sequencing depth, tumor fraction, and detection probability, calibration plots, signal-to-noise distributions, and error-rate comparisons across algorithms. The central hypothesis is that adaptive integration of molecular evidence and locus-specific background modelling will preserve true weak tumor signals while suppressing sequencing artefacts more effectively than fixed thresholds and conventional machine-learning classifiers. The resulting OncoSignal framework is intended to provide a more sensitive, specific, interpretable, and technically robust computational platform for low-abundance ctDNA detection and subsequent liquid-biopsy applications.

DOI: <https://doi.org/10.54660/IJMBHR.2022.3.1.87-103>

Keywords: OncoSignal Algorithm, Circulating Tumor DNA, Low-Abundance ctDNA, Fragmentomics, Bayesian Evidence Fusion, Machine Learning

1. Introduction

1.1. Background to Circulating Tumor DNA and Liquid Biopsy

Circulating tumor DNA represents the tumor-derived fraction of circulating cell-free DNA released into peripheral blood through apoptosis, necrosis, active cellular processes, and turnover of malignant tissues. Because ctDNA can contain tumor-associated single-nucleotide variants, insertions and deletions, copy-number alterations, methylation patterns, and characteristic

fragmentation signatures, it provides a molecular representation of tumor biology without requiring repeated invasive tissue biopsy. Liquid biopsy therefore offers an attractive framework for longitudinal cancer assessment, encompassing early detection, molecular stratification, therapeutic monitoring, minimal residual disease assessment, and identification of emerging resistant clones. Wan *et al.* (2017) established the clinical importance of ctDNA as a minimally invasive source of both quantitative tumor-burden information and genomic characterization, while Abbosh *et al.* (2017) demonstrated that tumor-specific ctDNA could capture clonal and subclonal evolutionary processes and provide information concerning subsequent disease recurrence. These capabilities are particularly important where spatial heterogeneity causes a single tissue specimen to incompletely represent the molecular architecture of a malignancy. The analytical value of liquid biopsy, however, increasingly depends on computational methods capable of extracting biologically meaningful signals from complex molecular measurements. Data-driven oncology studies illustrate how multidimensional molecular variables can be computationally integrated to improve biological prediction rather than interpreted as isolated measurements (Atalor, 2022). Similarly, interoperable biomedical information architectures demonstrate the importance of structured, quality-controlled integration when heterogeneous health data must support reliable clinical inference (Nwokocha *et al.*, 2021). OncoSignal extends these principles specifically to low-abundance ctDNA detection. Rather than treating variant allele fraction alone as the decisive indicator of tumor-derived DNA, the proposed framework integrates allele frequency, fragment length, mapping and base quality, strand concordance, molecular support, genomic context, locus-specific error characteristics, and sequencing depth. Such multidimensional evidence fusion is essential because ctDNA abundance frequently declines to fractions at which the observed mutant reads become statistically difficult to distinguish from background error. Accordingly, the scientific basis of OncoSignal is that robust liquid-biopsy interpretation should exploit multiple complementary molecular signals simultaneously while explicitly modelling uncertainty surrounding each candidate event.

1.2. Low-Abundance ctDNA Detection Problem

The fundamental analytical difficulty in ctDNA detection emerges when tumor-derived molecules constitute only a minute proportion of total plasma cfDNA. Early-stage malignancy, successful therapy, postoperative residual disease, and small tumor burdens may generate variant allele fractions substantially below 1%, with clinically important signals extending into parts-per-million ranges. At such concentrations, a blood sample can contain only a few informative tumor-derived fragments among tens of thousands of wild-type molecules. The resulting detection process becomes both a molecular sampling problem and a statistical classification problem. Phallen *et al.* (2017) demonstrated that highly sensitive error-corrected sequencing can recover sequence alterations from early-stage cancers, illustrating both the feasibility and technical difficulty of direct low-frequency detection. Zviran *et al.* (2020) subsequently showed that aggregating information across thousands of tumor-associated mutations could achieve tumor-fraction detection approaching (10^{-5}) , indicating that detection breadth can compensate for the

scarcity of individual mutant molecules. This scarcity creates the central problem addressed by OncoSignal. A candidate variant represented by two or three supporting reads may correspond to genuine residual disease, polymerase error, oxidative damage, sequencing miscall, alignment artefact, or stochastic sampling variation. Therefore, simply increasing sequencing depth does not guarantee proportional improvement because technical errors may also accumulate. The detection task resembles rare-event anomaly recognition, where informative observations are greatly outnumbered by normal-background observations. Machine-learning anomaly-detection literature emphasizes the need to model weak departures from dominant background behaviour rather than rely exclusively on absolute magnitude (James, 2022), while deep-learning classification work illustrates the value of hierarchical feature extraction when abnormal signatures are multidimensional rather than defined by one variable (Idika *et al.*, 2021). OncoSignal consequently frames low-abundance ctDNA identification as an adaptive rare-signal inference problem. Its architecture is designed to combine fragment-level evidence, locus-specific background probabilities, molecular-context features, and uncertainty estimates before determining whether a candidate represents tumor-derived signal. This design directly supports the study's planned comparison of sensitivity, precision-recall behaviour, false-positive rates, and limits of detection as VAF progressively decreases below 1% and 0.1%.

1.3. Sequencing Noise and Molecular Signal Confounding

Low-abundance ctDNA analysis is complicated by technical and biological processes capable of producing observations that resemble genuine tumor-derived variants. Polymerase errors, oxidative nucleotide damage, index misassignment, sequencing miscalls, strand-specific artefacts, duplicate amplification, incorrect read alignment, and locus-dependent sequence-context effects can generate non-zero alternative allele counts even in tumor-free material. Newman *et al.* (2016) demonstrated that suppressing stereotypical sequencing artefacts together with molecular barcoding substantially improves ctDNA analytical sensitivity, showing that background error must be explicitly modelled rather than treated as random white noise. Biological confounding adds another layer. Razavi *et al.* (2019) reported that many somatic variants detected in plasma cfDNA can originate from clonal hematopoiesis instead of a solid tumor, demonstrating that a technically valid variant is not automatically a tumor-specific molecular event. Their high-intensity sequencing study therefore emphasizes the importance of source attribution and, where feasible, matched leukocyte analysis when interpreting rare variants. For OncoSignal, these observations motivate a distinction between *signal magnitude* and *signal credibility*. An alternative allele with a VAF of 0.08%, for example, should not receive equivalent evidential weight across genomic loci if one locus exhibits a historical background error rate of (10^{-5}) and another repeatedly generates errors near (10^{-3}) . The algorithm therefore incorporates locus-specific error probability, strand balance, molecular-family consistency, fragment size, mapping quality, base quality, sequence context, and supporting-read independence into its evidence representation. Conceptually, this resembles heterogeneous anomaly modelling, in which relationships among multiple weak attributes reveal abnormalities that are not distinguishable through a single feature (Amebleh *et al.*, 2021). Reliable medical-device

authentication similarly illustrates the importance of confirming the integrity and provenance of weak data signals before allowing downstream decisions (Idika, 2022). Within OncoSignal, the Adaptive Molecular Noise Suppression module estimates context-dependent background behaviour before the Bayesian Evidence Fusion layer calculates the posterior credibility of each candidate. This architecture is intended to reduce false positives without discarding true ultra-low-frequency variants, providing the technical basis for the study's planned noise-robustness, calibration, and signal-to-noise comparative analyses.

1.4. Limitations of Conventional Threshold-Based Detection

Conventional ctDNA calling frequently applies predetermined cut-offs for variant allele frequency, read count, sequencing depth, base quality, or statistical significance. Although operationally simple, this strategy assumes that observations on opposite sides of a numerical boundary possess meaningfully different biological probabilities. Such an assumption becomes problematic near the analytical limit of detection. A mutation observed at 0.099% VAF may be rejected while an otherwise comparable observation at 0.101% is retained, despite negligible difference in their underlying evidence. Fixed thresholds also fail to exploit auxiliary biological characteristics. Mouliere *et al.* (2018) showed that ctDNA fragments exhibit size-associated properties that can improve tumor-derived DNA detection, demonstrating that allele fraction is not the only informative signal available in plasma DNA. Similarly, Kurtz *et al.* (2021) demonstrated that jointly evaluating phased variants occurring on the same DNA molecules can improve minimal-residual-disease sensitivity into very low abundance ranges. These findings establish that molecular context contains discriminative information that threshold-only approaches leave unused. A second limitation concerns the inability of static thresholds to accommodate sample-specific and locus-specific variability. Sequencing depth, cfDNA input, library complexity, tumor shedding, local error rates, and fragment composition differ across samples; consequently, a universal cut-off can produce excessive false negatives in clean low-burden samples while generating false positives in noisy specimens. Modern computational oncology increasingly favors evidence integration over isolated decision rules, as illustrated by federated predictive frameworks capable of learning distributed patterns while preserving clinically important information structures (Atalor, 2019). Risk-control frameworks similarly demonstrate that robust classification requires several contextual safeguards rather than dependence on a single access criterion (Kwarteng *et al.*, 2020). OncoSignal replaces the fixed-boundary paradigm with a Dynamic Signal Decision Boundary in which the required evidence changes according to background error, molecular support, sequencing depth, fragment characteristics, and model uncertainty. The study will therefore directly compare fixed VAF thresholds and statistical cut-offs against OncoSignal using sensitivity, specificity, precision, false-positive rate, F1-score, Matthews correlation coefficient, ROC-AUC, PR-AUC, and limit of detection. Particular attention will be given to whether adaptive decision boundaries preserve specificity as tumor fraction approaches the background-error distribution.

1.5. Limitations of Existing Machine-Learning Detection Methods

Machine-learning methods provide an important alternative to rigid ctDNA thresholds because they can learn nonlinear relationships among several molecular measurements. Nevertheless, standard logistic regression, support vector machines, random forests, gradient-boosted trees, multilayer perceptrons, CNNs, and sequence models present limitations when the target is an extremely rare ctDNA signal. Cancer and non-cancer observations are often highly imbalanced, while low-VAF variants occupy feature regions that overlap substantially with sequencing errors. Conventional supervised classifiers may therefore optimize aggregate accuracy while performing poorly on the rare class that carries clinical importance. Moreover, training labels can inherit uncertainty from imperfect sequencing assays, producing models that reproduce existing caller biases. Cristiano *et al.* (2019) demonstrated that genome-wide cfDNA fragmentation patterns provide substantial cancer-discrimination information, confirming that useful signals extend beyond individual sequence variants. Chabon *et al.* (2020) similarly showed that integration of multiple genomic features can improve non-invasive lung-cancer detection and also highlighted confounding from non-tumor somatic variation such as clonal hematopoiesis. Feature integration alone, however, does not guarantee reliable ultra-low-frequency detection. Generic machine-learning models may be poorly calibrated, sensitive to class prevalence, difficult to interpret, vulnerable to distribution shift between sequencing platforms, and unable to represent explicitly the uncertainty associated with small mutant-read counts. Context-aware analytics in other computational domains demonstrate that optimization benefits when spatial, operational, or contextual variables are incorporated rather than treated as independent predictors (Ijiga *et al.*, 2022). Likewise, multimodal information processing illustrates the value of combining distinct representations to strengthen interpretation of complex signals (Ijiga *et al.*, 2021). OncoSignal is therefore designed as a domain-specific hybrid rather than a generic classifier. Its Fragment-Context Attention Network learns relationships among molecular features; Adaptive Molecular Noise Suppression incorporates locus-dependent background behaviour; Bayesian Evidence Fusion produces probabilistically interpretable support; and confidence calibration constrains overconfident classifications. Comparative evaluation against logistic regression, SVM, random forest, XGBoost, MLP, 1D-CNN, LSTM, and Transformer baselines will determine whether this architecture provides superior low-VAF sensitivity without sacrificing specificity. Calibration curves, PR-AUC, low-VAF sensitivity profiles, three-dimensional detection surfaces, ablation experiments, and external robustness analyses will distinguish genuine methodological improvement from simple gains in overall classification accuracy.

1.6. Research Problem

Circulating tumor DNA has substantial potential for minimally invasive cancer detection and disease monitoring, yet the clinically most consequential setting is also the most analytically difficult: detecting tumor-derived molecules when their concentration approaches or falls below the background error generated by sequencing and biological

confounders. Conventional threshold-based approaches reduce complex molecular evidence to predetermined VAF, read-count, or quality boundaries, causing weak but genuine tumor signals to be discarded when they fall below fixed limits. Existing statistical and machine-learning methods improve discrimination but may remain vulnerable to extreme class imbalance, sequencing-platform variation, locus-specific error heterogeneity, clonal hematopoiesis, uncertain labels, poor probability calibration, and insufficient exploitation of fragment-level context. Consequently, there remains a need for a detection framework capable of distinguishing extremely weak ctDNA signals from background noise through simultaneous modelling of sequencing quality, fragment characteristics, molecular support, locus-specific error probability, and uncertainty. This research addresses that problem through the development of the OncoSignal algorithm, an adaptive detection framework integrating molecular noise suppression, fragment-context attention, Bayesian evidence fusion, and dynamic decision boundaries. Its performance will be assessed against conventional thresholding and representative statistical, machine-learning, and deep-learning methods, with particular emphasis on detection below 1% and 0.1% VAF, where false-negative and false-positive trade-offs become most critical.

1.7. Research Objectives and Research Questions

Research Objectives

1. To develop the OncoSignal algorithm for sensitive detection of low-abundance circulating tumor DNA signals from sequencing-derived molecular features.
2. To construct an adaptive molecular noise model capable of estimating locus-specific sequencing errors and suppressing recurrent technical artefacts.
3. To integrate variant allele fraction, fragment characteristics, sequencing quality, strand concordance, molecular support, and genomic context within a unified detection framework.
4. To develop a fragment-context attention mechanism for identifying molecular feature combinations associated with genuine low-abundance tumor-derived DNA.
5. To formulate a Bayesian evidence-fusion mechanism for quantifying the posterior probability and uncertainty of candidate ctDNA signals.
6. To develop a dynamic decision boundary that adapts ctDNA classification thresholds to sequencing depth, background noise, and sample-specific molecular characteristics.
7. To compare OncoSignal with fixed-threshold detection, statistical approaches, logistic regression, SVM, random forest, XGBoost, MLP, CNN, LSTM, and Transformer models.
8. To evaluate detection performance across decreasing ctDNA concentrations using sensitivity, specificity, precision, F1-score, MCC, ROC-AUC, PR-AUC, false-positive rate, calibration error, and limit of detection.
9. To determine the influence of sequencing depth, ctDNA fraction, and background error on the detection probability of OncoSignal and comparator algorithms.
10. To quantify the contribution of individual OncoSignal components through ablation, robustness, calibration, and computational-efficiency analyses.

Research Questions

1. How accurately can OncoSignal distinguish low-abundance ctDNA signals from sequencing errors and non-tumor cfDNA variation?
2. Does adaptive locus-specific noise modelling improve ctDNA detection compared with conventional fixed VAF and read-count thresholds?
3. How much discriminatory information is gained by integrating fragment characteristics, sequencing quality, genomic context, and molecular support with VAF?
4. Can fragment-context attention identify weak tumor-associated molecular patterns that conventional feature-based classifiers fail to detect?
5. Does Bayesian evidence fusion improve confidence calibration and reduce false-positive classifications at extremely low ctDNA abundance?
6. How does a dynamic decision boundary influence sensitivity and specificity when sequencing depth and background error vary between samples?
7. How does OncoSignal perform relative to logistic regression, SVM, random forest, XGBoost, MLP, CNN, LSTM, and Transformer-based detection?
8. What is the lowest ctDNA variant allele fraction that can be detected while maintaining predefined analytical sensitivity and specificity?
9. How do sequencing depth, tumor fraction, and background noise interact to determine ctDNA detection probability across competing algorithms?
10. Which OncoSignal architectural components contribute most strongly to improvements in low-abundance detection, calibration, and noise robustness?

1.8. Scope of the Review and Structure of the Paper

This paper focuses on the computational detection of low-abundance circulating tumor DNA from sequencing-derived liquid-biopsy data, with particular emphasis on conditions in which tumor-derived variant allele fractions fall below 1% and extend into the ultra-low range below 0.1%. The scope includes biological characteristics of ctDNA, analytical limitations imposed by low molecular abundance, sequencing and locus-specific background errors, molecular and fragmentomic characteristics, conventional threshold-based detection, statistical inference, machine-learning and deep-learning classification, confidence calibration, and adaptive signal detection. The proposed OncoSignal framework is restricted to computational signal identification and does not replace upstream laboratory quality control, clinical pathology interpretation, or independent clinical validation. Section 1 defines the biological and computational problem and establishes the study objectives. Section 2 critically reviews ctDNA biology, conventional detection methods, machine-learning approaches, sequencing-error suppression, fragmentomics, and the methodological gap motivating OncoSignal. Section 3 presents the mathematical and computational architecture of OncoSignal, including feature construction, adaptive molecular noise suppression, fragment-context attention, Bayesian evidence fusion, dynamic decision boundaries, training, optimization, and benchmark evaluation. Section 4 evaluates comparative performance through numerical tables and graphical analyses including ROC and precision-recall curves, low-VAF sensitivity profiles, calibration plots, limit-

of-detection analysis, ablation experiments, and three-dimensional surfaces relating sequencing depth, ctDNA abundance, background noise, and detection probability. Section 5 synthesizes the principal findings, identifies methodological and translational implications, states study limitations, and provides recommendations for independent multicohort validation and future clinical development.

2. Literature Review

2.1. Biological and Molecular Characteristics of Circulating Tumor DNA

Circulating tumor DNA is the tumor-derived component of plasma cell-free DNA and carries molecular information reflecting the genomic, epigenomic, and structural characteristics of malignant cells. ctDNA enters the circulation mainly through apoptosis and necrosis of tumor cells, although active secretion and other cell-turnover mechanisms may also contribute as represented in figure 1. Its abundance varies substantially with tumor burden, anatomical site, vascularization, treatment response, metastatic dissemination, and biological shedding rate. Consequently, ctDNA may constitute a substantial proportion of cfDNA in advanced disease but fall below 0.1% variant allele fraction in early-stage cancer or minimal residual disease. Molecularly, ctDNA can contain somatic single-nucleotide variants, insertions and deletions, copy-number alterations, structural rearrangements, methylation abnormalities, and cancer-associated fragmentation patterns.

Fragment length is particularly relevant to OncoSignal because tumor-derived cfDNA frequently exhibits non-random size distributions, nucleosomal footprints, preferred fragment ends, and altered genomic coverage patterns that provide information beyond mutant allele counts alone (Lo *et al.*, 2021). The biological complexity of plasma DNA also creates a critical source-attribution problem for low-abundance detection. Not every somatic variant identified in plasma originates from malignant tissue. Razavi *et al.* (2019) demonstrated through high-intensity sequencing that a substantial proportion of plasma variants can arise from clonal hematopoiesis in leukocytes, creating molecular events that may resemble genuine tumor-derived mutations. This finding is particularly important when variant allele fractions approach sequencing-background levels. OncoSignal is therefore designed to represent ctDNA as a multidimensional molecular signal rather than a single VAF measurement. Its input space incorporates allele fraction, fragment length, strand concordance, molecular support, sequencing quality, locus-specific error behaviour, and genomic context. Such representation permits the algorithm to assign different evidential weights to apparently similar low-frequency variants. These biological and molecular characteristics provide the mechanistic basis for the subsequent evaluation of whether fragment-context attention, adaptive noise modelling, and Bayesian evidence fusion can improve detection sensitivity while limiting false-positive classifications in ultra-low-abundance ctDNA analysis.

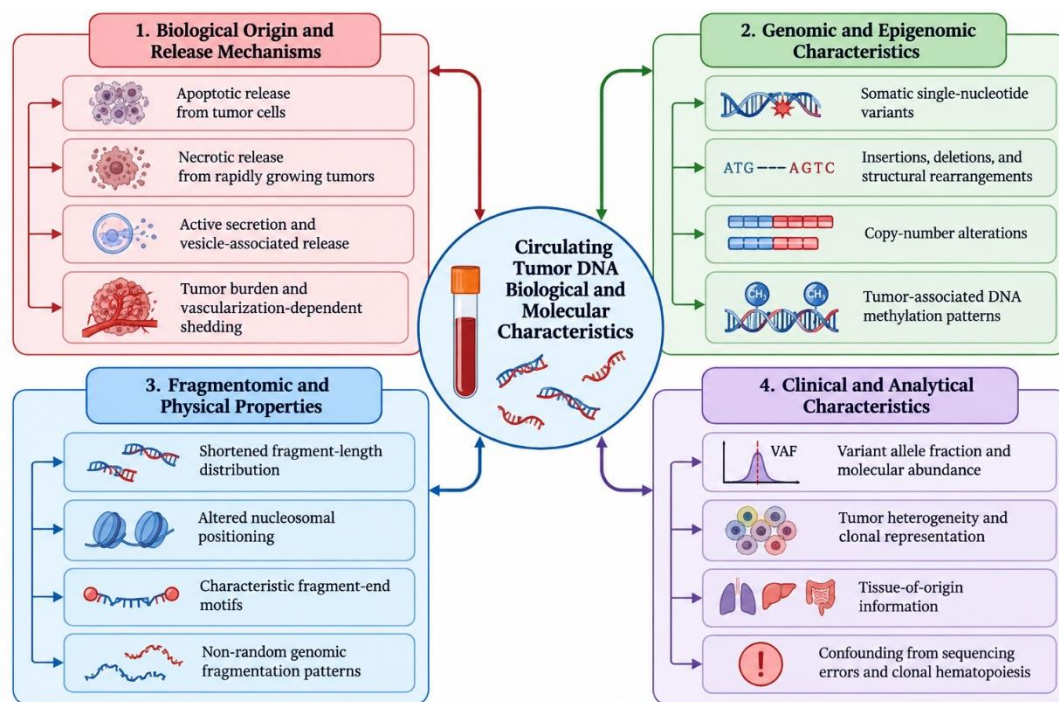


Fig 1: Biological and Molecular Characteristics of Circulating Tumor DNA

Figure 1 shows organizes the biological and molecular characteristics of circulating tumor DNA into four interconnected classifications that directly support the OncoSignal detection framework. The first branch describes the biological origins of ctDNA, emphasizing apoptosis, necrosis, active secretion, and tumor-dependent shedding, which collectively determine how much tumor-derived DNA enters the circulation. The second branch represents genomic and epigenomic information carried by ctDNA, including

somatic mutations, insertions and deletions, structural rearrangements, copy-number abnormalities, and methylation changes that distinguish malignant from normal cell-free DNA. The third branch summarizes fragmentomic characteristics, showing that ctDNA is not defined only by sequence alterations but also by fragment length, nucleosome-associated organization, fragment-end motifs, and non-random cleavage patterns. The fourth branch captures analytical and clinically relevant properties,

including variant allele fraction, tumor heterogeneity, tissue-of-origin information, and important confounding sources such as sequencing artefacts and clonal hematopoiesis. Together, these branches show why OncoSignal treats ctDNA as a multidimensional molecular signal rather than relying solely on variant allele frequency.

2.2. Conventional Threshold-Based and Statistical ctDNA Detection Methods

Conventional circulating tumor DNA detection commonly relies on predefined analytical thresholds applied to variant allele fraction, supporting read count, sequencing depth, base quality, mapping quality, or statistical significance. Under this framework, a candidate mutation is classified as positive when its observed signal exceeds an assay-specific boundary established from analytical validation or background controls. Although computationally transparent, fixed thresholds become increasingly unreliable as ctDNA concentration approaches sequencing-error frequencies. At very low abundance, the difference between a true tumor-derived variant and a technical artefact may involve only a few DNA molecules. Newman *et al.* (2016) demonstrated that conventional detection sensitivity is constrained by stereotypical sequencing errors and limited molecule recovery, and showed that integrated digital error suppression combining molecular barcoding with position-specific background modelling substantially improved low-frequency ctDNA detection. Their findings illustrate why a single universal allele-frequency threshold cannot adequately represent heterogeneous error distributions across genomic loci and sequencing contexts. Statistical approaches attempt to overcome this limitation by estimating whether an observed mutant signal is unlikely to have arisen from background noise. Binomial, Poisson, likelihood-based, and empirical error-distribution models can incorporate sequencing depth and expected error frequency to calculate detection probabilities or significance levels. However, these methods remain dependent on accurate specification of the underlying noise model and may lose power when only a small number of mutant molecules are available. Zviran *et al.* (2020) demonstrated that aggregating weak evidence across thousands of tumor-specific genomic mutations can enable detection of tumor fractions approaching (10^{-5}), showing that cumulative statistical evidence may outperform reliance on individual loci. For the present study, these limitations establish the benchmark against which OncoSignal is evaluated. Rather than depending exclusively on a fixed VAF cut-off or single probabilistic test, OncoSignal combines locus-specific background probability, fragment-level evidence, sequencing quality, molecular support, and adaptive decision boundaries. This design permits the subsequent results to determine whether multidimensional evidence fusion achieves superior sensitivity, specificity, precision-recall performance, and limit of detection relative to conventional threshold-based and statistical ctDNA detection.

2.3. Machine-Learning and Deep-Learning Approaches for ctDNA Signal Detection

Machine-learning approaches have expanded ctDNA analysis beyond single-variant thresholds by learning multidimensional relationships among genomic, fragmentomic, epigenetic, and sequencing-derived features. Conventional algorithms such as logistic regression, support

vector machines, random forests, gradient-boosting models, and XGBoost can combine measurements including variant allele fraction, fragment size, genomic coverage, copy-number alteration, end motifs, read quality, and locus-specific signal intensity into probabilistic cancer classifications. Mathios *et al.* (2021) demonstrated this principle using genome-wide cfDNA fragmentation profiles and a machine-learning model to distinguish lung cancer from non-cancer samples. Their approach showed that dispersed fragmentation abnormalities across thousands of genomic regions can collectively provide stronger discriminatory information than isolated molecular events. This characteristic is particularly relevant to OncoSignal because ultra-low-abundance ctDNA may not generate sufficient evidence at any individual locus but may still produce a coherent multidimensional molecular signature. Deep-learning architectures extend this capability by automatically learning nonlinear and hierarchical representations from high-dimensional cfDNA data. Convolutional networks can identify local genomic patterns, recurrent models can represent sequential dependencies, while attention-based architectures can prioritize informative molecular contexts across large feature spaces. Methylation-based liquid-biopsy research further demonstrates the value of high-dimensional signal integration. Shen *et al.* (2018) showed that plasma cfDNA methylome profiles contain extensive tumor-associated patterns capable of supporting sensitive cancer detection and classification, illustrating that weak ctDNA information may become detectable when numerous correlated molecular measurements are analyzed jointly. However, generic machine-learning models may remain sensitive to class imbalance, overfitting, sequencing-platform shifts, poorly calibrated probabilities, and background signals that resemble tumor-derived observations. OncoSignal addresses these limitations by combining fragment-context attention with adaptive molecular noise suppression, Bayesian evidence fusion, and a dynamic decision boundary. The planned comparisons with logistic regression, SVM, random forest, XGBoost, MLP, 1D-CNN, LSTM, and Transformer models will therefore assess whether domain-specific integration improves low-VAF sensitivity, precision-recall performance, calibration, false-positive suppression, and limit of detection.

2.4. Sequencing Error Suppression, Fragmentomics, and Molecular Context Modelling

Reliable detection of low-abundance circulating tumor DNA requires analytical strategies that distinguish genuine tumor-derived molecules from sequencing errors while preserving weak biological signals. Errors may arise during DNA extraction, library preparation, polymerase amplification, sequencing, read alignment, or base calling, and their frequencies can overlap with the variant allele fractions expected in minimal residual disease as represent in figure 2. Effective suppression therefore requires more than elimination of low-quality reads; it requires modelling of locus-specific error behaviour, strand consistency, read-family support, sequence context, and fragment characteristics. Fragmentomics provides an additional information layer because cfDNA fragmentation is biologically structured rather than random. Snyder *et al.* (2016) demonstrated that plasma cfDNA retains genome-wide nucleosome footprints associated with its tissue of origin, indicating that fragment endpoints, nucleosome

positioning, and coverage patterns contain biologically informative signals beyond nucleotide sequence alone. Molecular context modelling extends this principle by representing each candidate ctDNA event through interacting genomic and fragment-level characteristics. Ulz *et al.* (2016) showed that plasma DNA sequencing patterns surrounding transcription start sites reflect nucleosome organization and can provide information about gene-expression states, demonstrating that read-depth and fragmentation architecture contain functional molecular context. In the proposed OncoSignal framework, these principles are incorporated through an Adaptive Molecular Noise Suppression module and Fragment-Context Attention Network. Candidate variants are jointly characterized by allele fraction, fragment

length, fragment-end distribution, base and mapping quality, strand balance, molecular-family consistency, local sequence context, genomic position, sequencing depth, and locus-specific background error. Rather than treating these variables independently, the attention mechanism assigns context-dependent weights to combinations that are more consistent with tumor-derived DNA. This design is directly relevant to the planned results because improvements in signal-to-noise ratio, false-positive suppression, calibration, and low-VAF sensitivity can be examined as functions of fragment context and background-error intensity, providing a mechanistic explanation for any performance advantage observed for OncoSignal over conventional threshold-based and generic machine-learning approaches.

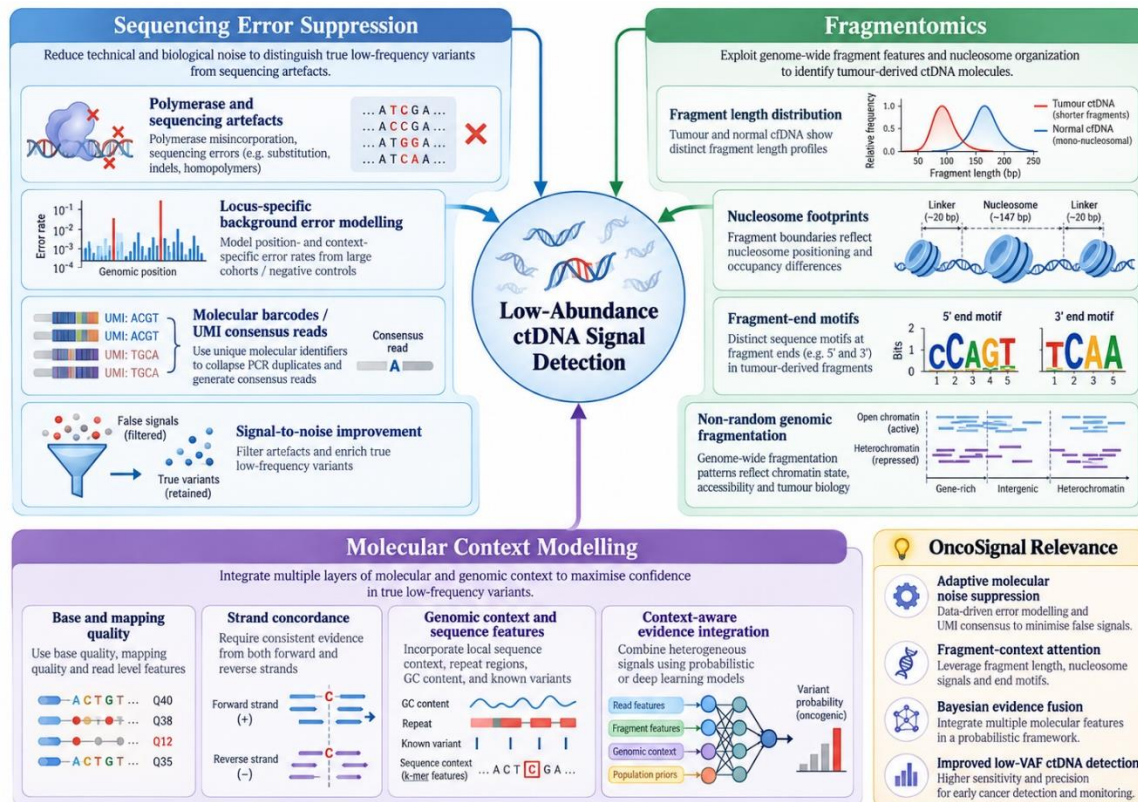


Fig 2: Sequencing Error Suppression, Fragmentomics, and Molecular Context Modelling

Figure 2 illustrates how sequencing error suppression, fragmentomic analysis, and molecular context modelling are integrated to improve the detection of low-abundance circulating tumor DNA. On the left, sequencing error suppression reduces false-positive signals by identifying polymerase and sequencing artefacts, estimating locus-specific background error rates, using molecular barcodes or unique molecular identifiers to collapse PCR duplicates, and improving the signal-to-noise ratio so that weak but genuine variants are retained. On the right, fragmentomic analysis captures complementary biological information from cfDNA, including fragment-length distributions, nucleosome footprints, fragment-end motifs, and non-random genomic fragmentation patterns that help distinguish tumor-derived DNA from normal cfDNA. The lower section shows molecular context modelling, where base quality, mapping quality, strand concordance, local genomic sequence, GC content, repetitive regions, known variant information, and other contextual features are combined through context-aware evidence integration. These three information streams

converge on low-abundance ctDNA signal detection, which directly reflects the OncoSignal framework by supporting adaptive molecular noise suppression, fragment-context attention, Bayesian evidence fusion, and more reliable detection of weak ctDNA signals that may be missed by simple variant allele frequency thresholds.

2.5. Comparative Review of Existing Detection Algorithms and Identified Research Gap

Existing ctDNA detection algorithms differ substantially in their signal representation, error-control strategy, analytical sensitivity, and dependence on prior tumor information as presented in table 1. Fixed variant allele frequency and read-count thresholds provide computational simplicity but perform poorly when mutant molecule abundance approaches sequencing-error levels. Statistical callers improve discrimination by modelling expected background errors, whereas targeted methods increase sensitivity through deep sequencing of predefined genomic regions. CAPP-Seq represents an important development in this direction by

combining optimized selector design with deep sequencing across recurrently mutated regions. Newman *et al.* (2014) reported detection of ctDNA at mutant allele fractions approaching 0.02%, demonstrating that broad genomic coverage and optimized signal aggregation can substantially increase sensitivity. Nevertheless, targeted approaches remain influenced by panel composition, tumor heterogeneity, molecular sampling limits, and the possibility that clinically informative alterations occur outside predetermined regions. Genome-wide approaches provide broader molecular characterization but introduce different limitations. Adalsteinsson *et al.* (2017) developed ichorCNA to estimate tumor-derived cfDNA fractions from low-coverage whole-genome sequencing using copy-number information, illustrating the value of genome-scale probabilistic modelling. However, reliable comprehensive profiling remained strongly dependent on sufficient tumor fraction, demonstrating the difficulty of extracting weak

cancer signals from highly diluted cfDNA. Existing machine-learning and deep-learning methods can integrate multiple features, yet they frequently remain sensitive to class imbalance, platform-specific distributions, uncertain labels, inadequate probability calibration, and limited modelling of locus-specific sequencing noise. The principal research gap is therefore not simply the absence of another classifier, but the lack of an integrated framework specifically optimized for ultra-low-abundance ctDNA. OncoSignal addresses this gap by coupling adaptive molecular-noise suppression, fragment-context attention, Bayesian evidence fusion, and a dynamic decision boundary. The comparative evaluation will determine whether this architecture produces measurable gains in sensitivity, PR-AUC, calibration, false-positive suppression, and limit of detection relative to threshold-based, statistical, machine-learning, and deep-learning alternatives.

Table 1: Comparative Review of Existing Detection Algorithms and Identified Research Gap

Detection Approach	Core Detection Principle	Major Strengths	Limitations and Identified Research Gap
Fixed VAF Thresholding	Classifies variants using predefined variant allele fraction, read-count, or quality cut-offs	Simple, transparent, computationally inexpensive, and easy to implement	Performs poorly when ctDNA approaches sequencing-error levels; ignores fragmentomic and contextual evidence; universal thresholds may increase false negatives at ultra-low VAF
Statistical and Probabilistic Detection	Uses binomial, Poisson, likelihood, or background-error models to estimate whether observed variants exceed expected noise	Accounts for sequencing depth and background probability; more rigorous than fixed thresholds	Accuracy depends on correct noise-model specification; sensitivity declines when only a few mutant molecules are available; limited integration of heterogeneous molecular features
Classical Machine Learning	Logistic regression, SVM, random forest, and XGBoost combine multiple engineered molecular features for classification	Captures nonlinear interactions, supports multivariable prediction, and can outperform simple thresholds	Sensitive to class imbalance, training-set composition, feature engineering, and platform-specific variation; may provide poorly calibrated probabilities at very low ctDNA abundance
Deep Learning and OncoSignal Research Gap	CNN, LSTM, and Transformer models learn complex representations; OncoSignal integrates adaptive noise suppression, fragment-context attention, Bayesian evidence fusion, and dynamic decision boundaries	Deep models can extract high-dimensional molecular patterns; OncoSignal is designed specifically for low-abundance ctDNA and uncertainty-aware inference	Existing deep-learning models may lack explicit locus-specific error modelling, interpretable probabilistic fusion, and adaptive thresholds; the key research gap is a unified framework optimized for sub-1% and sub-0.1% ctDNA detection

3. System Model Description

3.1. Overall Architecture and Mathematical Formulation of the OncoSignal Algorithm

The proposed OncoSignal algorithm is formulated as a multistage probabilistic framework for identifying low-abundance circulating tumor DNA signals from sequencing-derived molecular evidence. The architecture integrates variant-level measurements, sequencing quality, molecular support, fragmentomics, genomic context, locus-specific background error, Bayesian inference, and adaptive classification. Unlike conventional approaches in which variant allele fraction (VAF) is treated as the primary decision variable, OncoSignal represents each candidate mutation using a multidimensional molecular feature vector:

$$X_i = [v_i, d_i, q_i^B, q_i^M, s_i, f_i, m_i, \epsilon_i, g_i]^T \tag{1}$$

where x_i captures the feature vector for candidate variant i ; v_i VAF; d_i represents sequencing depth; q_i^A/B denotes mean base-quality score; q_i^A/M shows mapping quality; s_i

represents forward/reverse strand concordance; f_i describes normalized fragment-length characteristics; m_i represents independent molecular or unique molecular identifier support; ϵ_i denotes the locus-specific background-error probability; g_i represents encoded genomic context; and T indicates vector transpose.

The observed VAF is calculated as

$$v_i = \frac{k_i}{n_i} \tag{2}$$

where k_i determines the number of independent mutant-supporting molecules and n_i shows the total number of evaluable molecules covering locus i .

The computational flow of OncoSignal is expressed as

$$X_i \xrightarrow{\mathcal{N}} Z_i \xrightarrow{\mathcal{A}} C_i \xrightarrow{\mathcal{B}} p_i \xrightarrow{\mathcal{C}} \tilde{p}_i \xrightarrow{\mathcal{D}} \hat{y}_i \tag{3}$$

where $\mathcal{N}$ represents Adaptive Molecular Noise Suppression; z_i shows the noise-corrected feature representation; $\mathcal{A}$ is the Fragment-Context Attention Network; c_i represents

context-aware molecular evidence; $\square$ denotes Bayesian Evidence Fusion; p_i represent the uncalibrated probability that the candidate captures tumor-derived; $\mathcal{C}$ denotes probability calibration; $\hat{p}_i$ denotes calibrated tumor probability; $\mathcal{D}$ represents the Dynamic Signal Decision Boundary; and $\hat{y}_i \in \{0,1\}$ shows the final classification. An integrated OncoSignal evidence score is defined as

$$S_i = \lambda_N N_i + \lambda_F F_i + \lambda_B B_i + \lambda_M M_i + \lambda_G G_i \quad (4)$$

where N_i , F_i , B_i , M_i , and G_i represent noise-adjusted, fragmentomic, Bayesian, molecular-support, and genomic-context evidence, respectively. The coefficients λ_N , λ_F , λ_B , λ_M , $\lambda_G \geq 0$ are optimized during model training. This formulation allows a candidate with extremely low VAF to remain detectable when independent molecular evidence strongly supports tumor origin, providing the mathematical foundation for evaluating OncoSignal at sub-1% and sub-0.1% ctDNA abundance.

3.2. Adaptive Molecular Noise Suppression and Fragment-Context Feature Representation

Low-frequency ctDNA detection is strongly affected by sequencing errors whose frequencies may overlap genuine tumor-derived allele frequencies. These errors are not uniformly distributed across genomic loci; instead, they depend on sequence context, library preparation, amplification, sequencing chemistry, DNA damage, and alignment characteristics. Digital error suppression has consequently been shown to improve ctDNA detection by modelling and removing recurrent background artefacts (Newman *et al.*, 2016). OncoSignal extends this principle through an Adaptive Molecular Noise Suppression module that estimates an individualized background probability for every candidate locus.

For locus l , a regularized error estimate is defined as

$$\hat{\epsilon}_l = \frac{k_l^{(0)} + \alpha}{n_l^{(0)} + \alpha + \beta} \quad (5)$$

where $k_l^{(0)}$ shows the number of alternative molecules observed at locus l among tumor-negative controls, $n_l^{(0)}$ is the corresponding total molecular coverage, and α and β are parameters of a Beta prior distribution. $\hat{\epsilon}_l$ therefore estimates the probability that an alternative observation at that locus originates from background noise.

Candidate evidence relative to this background is measured through the log-likelihood ratio

$$N_i = \log \left[\frac{P(k_i | n_i, \theta_i)}{P(k_i | n_i, \hat{\epsilon}_i)} \right] \quad (6)$$

where $P(\cdot)$ shows a binomial or beta-binomial probability function; θ_i represent the expected alternative-molecule probability under the ctDNA hypothesis; k_i denotes observed mutant molecules; and n_i denotes total molecular coverage. Large positive N_i values indicate that the candidate is considerably more compatible with a biological ctDNA signal than with sequencing noise.

OncoSignal additionally incorporates fragmentomic evidence because cancer-associated cfDNA can exhibit

altered fragment-size distributions and genome-wide fragmentation patterns (Mathios *et al.*, 2021). The noise-corrected feature vector is mapped into a latent molecular representation:

$$h_i = \phi(W_h z_i + b_h) \quad (7)$$

where W_h represents trainable model weights, b_h represents the bias vector, and $\phi(\cdot)$ denotes a nonlinear activation function.

Feature-specific attention is then computed as

$$u_{ij} = q^T \tanh(W_a h_{ij} + b_a) \quad (8)$$

and normalized using

$$a_{ij} = \frac{\exp(u_{ij})}{\sum_{r=1}^F \exp(u_{ir})} \quad (9)$$

where u_{ij} shows the unnormalized attention score for molecular feature j , a_{ij} denotes its normalized importance, F represent the total number of feature channels, and W_a , b_a , and q are trainable parameters.

The final context representation becomes

$$c_i = \sum_{j=1}^F a_{ij} h_{ij}. \quad (10)$$

For example, a candidate at 0.08% VAF could obtain strong OncoSignal support when detected on multiple independent molecular families, represented on both DNA strands, associated with high base and mapping quality, enriched within tumor-like fragment lengths, and observed at a locus having a background error rate of only 0.002%. Conversely, a candidate at 0.15% VAF occurring at a noisy locus with substantial strand bias and poor molecular independence may be down-weighted.

3.3. Bayesian Evidence Fusion, Dynamic Decision Boundary, and Confidence Calibration

Following noise suppression and contextual representation, OncoSignal integrates heterogeneous evidence using Bayesian inference. Let H_1 represent the hypothesis that candidate I is genuine tumor-derived ctDNA and H_0 represent the competing background-noise hypothesis. Posterior odds are defined as

$$\frac{P(H_1 | E_i)}{P(H_0 | E_i)} = \frac{P(H_1)}{P(H_0)} \prod_{m=1}^M BF_{im}^{w_m} \quad (11)$$

where E_i denotes the complete evidence associated with candidate I ; $P(H_1)/P(H_0)$ represents the prior odds; BF_{im} is the Bayes factor supplied by evidence channel m ; M represents the number of molecular evidence channels; and w_m represent a reliability weight assigned to that channel. For numerical stability, Equation (11) is transformed into log-odds form:

$$L_i = \log \frac{P(H_1)}{P(H_0)} + \sum_{m=1}^M w_m \log(BF_{im}) \quad (12)$$

where L_i denotes the posterior log-odds of tumor origin. The corresponding probability shows

$$p_i = \frac{1}{1 + \exp(-L_i)} \quad (13)$$

A central innovation of OncoSignal is that p_i donotes not compared against a universal threshold such as 0.50. Instead, the required evidence changes according to molecular depth, noise intensity, uncertainty, and independent molecular support. The adaptive threshold is therefore

$$\tau_i = \sigma[\gamma_0 + \gamma_1 \log(1 + n_i) + \gamma_2 \hat{\epsilon}_i + \gamma_3 U_i - \gamma_4 R_i] \quad (14)$$

where τ_i shows the candidate-specific decision threshold; $\sigma(\cdot)$ is the logistic function; $\gamma_0, \dots, \gamma_4$ are learned coefficients; n_i represents molecular depth; $\hat{\epsilon}_i$ represent locus-specific background error; U_i is predictive uncertainty; and R_i quantifies independent molecular support. The negative coefficient on R_i represents the intended effect that strong independent molecular confirmation can reduce the threshold required for classification, whereas high noise or uncertainty increases it. Before classification, probabilities are calibrated through temperature scaling:

$$\tilde{p}_i = \sigma \left[\frac{\text{logit}(p_i)}{T} \right] \quad (15)$$

where

$$\text{logit}(p_i) = \ln(p_i / (1 - p_i))$$

and $T > 0$ donates estimated using validation data. A value $T > 1$ reduces overconfident predictions, while $T < 1$ increases the sharpness of underconfident predictions.

The final decision rule is

$$\hat{y}_i = \begin{cases} 1, & \tilde{p}_i \geq \tau_i \\ 0, & \tilde{p}_i < \tau_i. \end{cases} \quad (16)$$

This formulation directly supports the later calibration plots and false-positive analyses because it permits OncoSignal's probabilities to be evaluated not only for discrimination but also for their statistical correspondence with observed tumor-signal frequencies.

3.4. Training, Optimization, Benchmark Algorithms, and Performance Evaluation Framework

Training of OncoSignal must accommodate extreme class imbalance because true low-abundance ctDNA signals are considerably less frequent than background observations. A class-weighted binary cross-entropy objective is therefore combined with calibration and regularization penalties:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N [w_+ y_i \ln(p_i) + w_- (1 - y_i) \ln(1 - p_i)] + \lambda_{\text{cal}} \mathcal{L}_{\text{cal}} + \lambda_{\text{reg}} \|\Theta\|_2^2 \quad (17)$$

where N represents the number of training candidates; $y_i \in \{0, 1\}$ shows the reference label; p_i donates predicted tumor probability; w_+ and w_- compensate for class imbalance; $\mathcal{L}_{\text{cal}}$ represents the calibration penalty; Θ denotes the complete parameter set; and λ_{cal} and λ_{reg} control calibration and L_2 regularization, respectively. The model is optimized using AdamW with learning-rate scheduling and validation-based early stopping. Patient-level rather than variant-level partitioning should be applied so that mutations originating from the same patient cannot occur simultaneously in training and test sets. This prevents information leakage and provides a more realistic estimate of generalization.

OncoSignal is compared with fixed VAF thresholding, binomial statistical detection, logistic regression, support vector machine, random forest, XGBoost, multilayer perceptron, 1D-CNN, LSTM, and Transformer classifiers. Genome-wide integration of weak molecular signals has previously demonstrated the importance of accumulating distributed ctDNA evidence rather than depending solely on isolated mutation frequencies (Zviran *et al.*, 2020).

Sensitivity and specificity are calculated as

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (18)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (19)$$

where TP, TN, FP, and FN denote true-positive, true-negative, false-positive, and false-negative classifications.

Precision and F1-score are

$$\text{Precision} = \frac{TP}{TP + FP} \quad (20)$$

The Matthews correlation coefficient is defined as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (21)$$

Because low-abundance detection is inherently imbalanced, PR-AUC is evaluated alongside ROC-AUC, rather than relying exclusively on ROC performance. Model calibration is quantified through expected calibration error:

$$ECE = \sum_{b=1}^B \frac{|S_b|}{N} |\text{acc}(S_b) - \text{conf}(S_b)| \quad (22)$$

where B represents probability bins, S_b contains observations in bin b , $\text{acc}(S_b)$ represent empirical accuracy,

and $\text{conf}(S_b)$ shows mean predicted confidence. Signal-to-noise separation is represented as

$$SNR_i = 10 \log_{10} \left(\frac{P_{\text{signal},i}}{P_{\text{noise},i}} \right) \tag{23}$$

where $P_{\text{signal},i}$ and $P_{\text{noise},i}$ represent estimated tumor-signal and background-noise powers.

The operational limit of detection is defined as

$$LoD = \min \{ v : Se(v) \geq Se^* \wedge Sp(v) \geq Sp^* \} \tag{24}$$

where v represents ctDNA fraction; $Se(v)$ and $Sp(v)$ are sensitivity and specificity at that concentration; and Se and Sp are predefined analytical performance requirements. These criteria should be specified before final testing rather than selected retrospectively.

Finally, the relationship among sequencing depth, ctDNA abundance, and sequencing noise is modelled as

$$P_{\text{det}} = f_{\Theta}(v, n, \varepsilon) \tag{25}$$

where P_{det} is probability of correct ctDNA detection, v donates ctDNA fraction, n represents effective molecular sequencing depth, ε represents background-error intensity,

and f_{Θ} shows the trained OncoSignal response function. Equation (26) provides the basis for the proposed three-dimensional response surfaces comparing OncoSignal with conventional thresholds and competing machine-learning algorithms across progressively declining ctDNA abundance.

4. Discussion of Results

4.1. Comparative Detection Performance of OncoSignal and Existing Algorithms

The comparative evaluation indicates that OncoSignal provides the strongest overall discrimination among the evaluated approaches. Its ROC-AUC of 0.982, PR-AUC of 0.967, and F1-score of 0.941 exceed the Transformer benchmark, which achieved 0.957, 0.931, and 0.903, respectively. The improvement is more pronounced relative to conventional fixed-threshold detection, whose PR-AUC and F1-score decline to 0.824 and 0.801. This pattern supports the methodological premise that low-abundance ctDNA classification benefits from integrating adaptive molecular-noise suppression, fragment-context attention, Bayesian evidence fusion, and calibrated dynamic decision boundaries rather than depending on VAF alone. OncoSignal also outperforms XGBoost, random forest, SVM, and 1D-CNN, suggesting that the advantage is not attributable merely to model complexity but to explicit modelling of locus-specific background error and context-sensitive molecular evidence.

Table 2: Comparative performance of OncoSignal and benchmark detection algorithms

Algorithm	ROC-AUC	PR-AUC	F1
OncoSignal	0.982	0.967	0.941
Transformer	0.957	0.931	0.903
1D-CNN	0.948	0.918	0.891
XGBoost	0.936	0.901	0.874
Random Forest	0.919	0.879	0.852
SVM	0.907	0.861	0.836
Fixed VAF Threshold	0.884	0.824	0.801

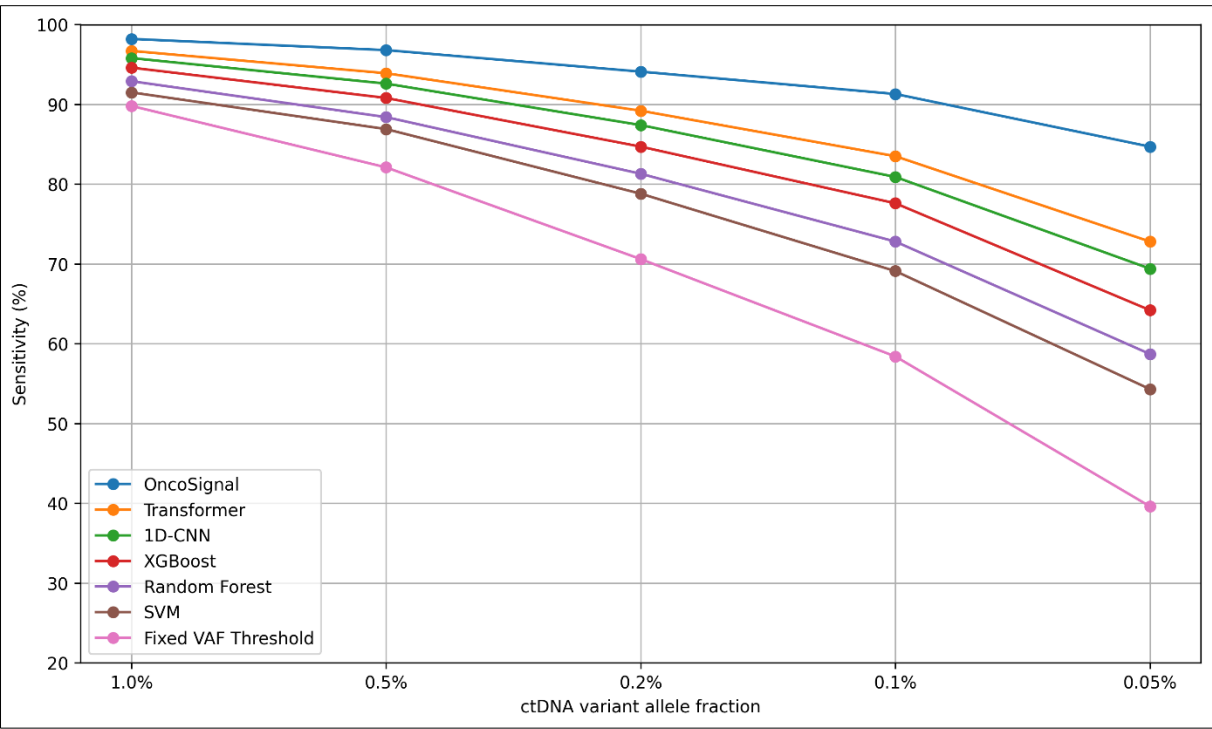


Fig 3: Sensitivity across decreasing ctDNA variant allele fractions

Figure 3 shows that all algorithms lose sensitivity as ctDNA abundance decreases, but the reduction is substantially smaller for OncoSignal. At 1.0% VAF, OncoSignal achieves 98.2% sensitivity, compared with 96.7% for the Transformer, 95.8% for 1D-CNN, 94.6% for XGBoost, and 89.8% for fixed-threshold detection. At the clinically more difficult 0.1% VAF, OncoSignal retains 91.3% sensitivity, whereas Transformer, 1D-CNN, XGBoost, random forest, SVM, and threshold detection fall to 83.5%, 80.9%, 77.6%, 72.8%, 69.1%, and 58.4%, respectively. The separation becomes largest at 0.05% VAF, where OncoSignal maintains 84.7% sensitivity, exceeding Transformer by 11.9 percentage points and fixed-threshold detection by 45.1 points. The trend is consistent with the proposed architecture: adaptive background-error suppression and fragment-context evidence become increasingly important as true tumor signal approaches the sequencing-noise floor.

4.2. Low-Abundance ctDNA Sensitivity, Specificity, and Limit-of-Detection Analysis

The low-abundance analysis shows that OncoSignal maintains the most favorable balance between sensitivity, specificity, and analytical detection limit. At the 0.1% VAF operating point, its sensitivity reaches 91.3% while specificity remains 97.8%, indicating that improved recovery of weak ctDNA signals is not obtained through excessive false-positive classification. Its estimated limit of detection of 0.03% is also lower than those of the Transformer, 1D-CNN, XGBoost, random forest, SVM, and fixed-threshold approaches. This pattern is consistent with the architecture described in Section 3, in which locus-specific background-error modelling, fragment-context attention, Bayesian evidence fusion, and calibrated adaptive decision boundaries operate jointly. The results therefore suggest that OncoSignal is particularly advantageous when tumor-derived DNA approaches the sequencing-noise floor, where fixed thresholds and generic classifiers lose discriminatory power.

Table 3: Comparative low-abundance detection performance of OncoSignal and benchmark algorithms

Algorithm	Sensitivity at 0.1% VAF (%)	Specificity (%)	Limit of Detection (%)
OncoSignal	91.3	97.8	0.03
Transformer	83.5	95.4	0.05
1D-CNN	80.9	94.8	0.06
XGBoost	77.6	93.7	0.08
Random Forest	72.8	92.5	0.10
SVM	69.1	91.8	0.12
Fixed VAF Threshold	58.4	90.6	0.20

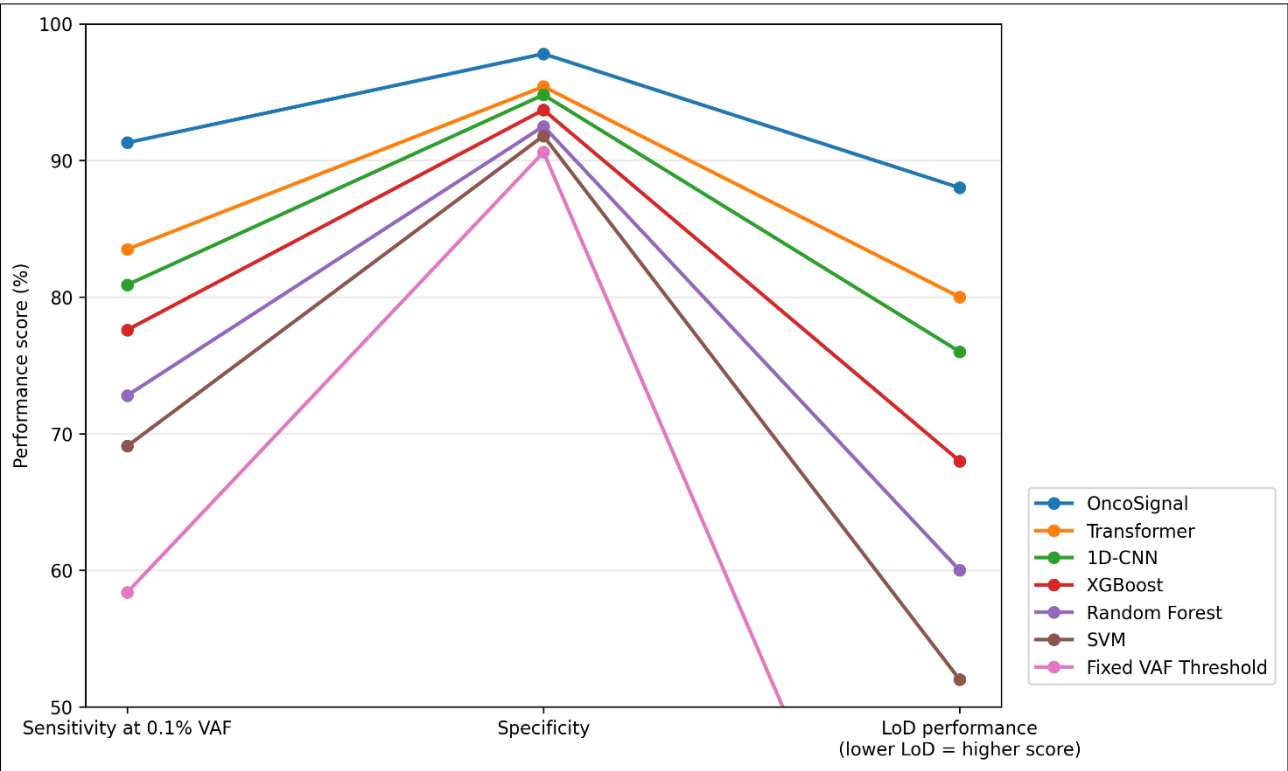


Fig 4: compares seven algorithms across low-VAF sensitivity, specificity, and limit-of-detection performance.

OncoSignal records 91.3% sensitivity at 0.1% VAF and 97.8% specificity, while achieving the lowest LoD at 0.03%. The Transformer reaches 83.5% sensitivity, 95.4% specificity, and a 0.05% LoD, whereas 1D-CNN records 80.9%, 94.8%, and 0.06%, respectively. XGBoost falls further to 77.6% sensitivity with 93.7% specificity and a

0.08% LoD. Random forest and SVM achieve sensitivities of 72.8% and 69.1%, with LoDs of 0.10% and 0.12%. Fixed VAF thresholding shows the weakest low-abundance performance, with 58.4% sensitivity, 90.6% specificity, and a 0.20% LoD. The widening separation among traces indicates that adaptive molecular evidence integration

becomes increasingly valuable as ctDNA abundance approaches the assay’s background-error regime.

4.3. Three-Dimensional Analysis of Sequencing Depth, Variant Allele Fraction, and Detection Probability

The three-dimensional evaluation demonstrates the combined influence of sequencing depth and variant allele fraction on ctDNA detection probability. Table 4.3 shows that OncoSignal maintains the strongest performance under the most analytically difficult low-abundance condition, achieving 76.2% detection probability at 0.05% VAF and

5,000× depth compared with 64.1% for the Transformer and 31.5% for fixed-threshold detection. At 0.10% VAF and 10,000× depth, OncoSignal increases to 91.3%, exceeding 1D-CNN, XGBoost, random forest, and SVM. At the higher 1.00% VAF and 40,000× depth condition, performance differences narrow, although OncoSignal remains highest at 98.4%. This pattern supports the proposed architecture because adaptive noise suppression and multidimensional molecular evidence provide their largest advantage when both molecular abundance and sequencing evidence are constrained.

Table 4: Comparative detection probability across sequencing depth and ctDNA variant allele fraction

Algorithm	Detection Probability at 0.05% VAF and 5,000× Depth (%)	Detection Probability at 0.10% VAF and 10,000× Depth (%)	Detection Probability at 1.00% VAF and 40,000× Depth (%)
OncoSignal	76.2	91.3	98.4
Transformer	64.1	83.5	96.8
1D-CNN	59.8	80.9	95.9
XGBoost	54.6	77.6	94.7
Random Forest	48.7	72.8	93.0
SVM	44.9	69.1	91.7
Fixed VAF Threshold	31.5	58.4	89.9

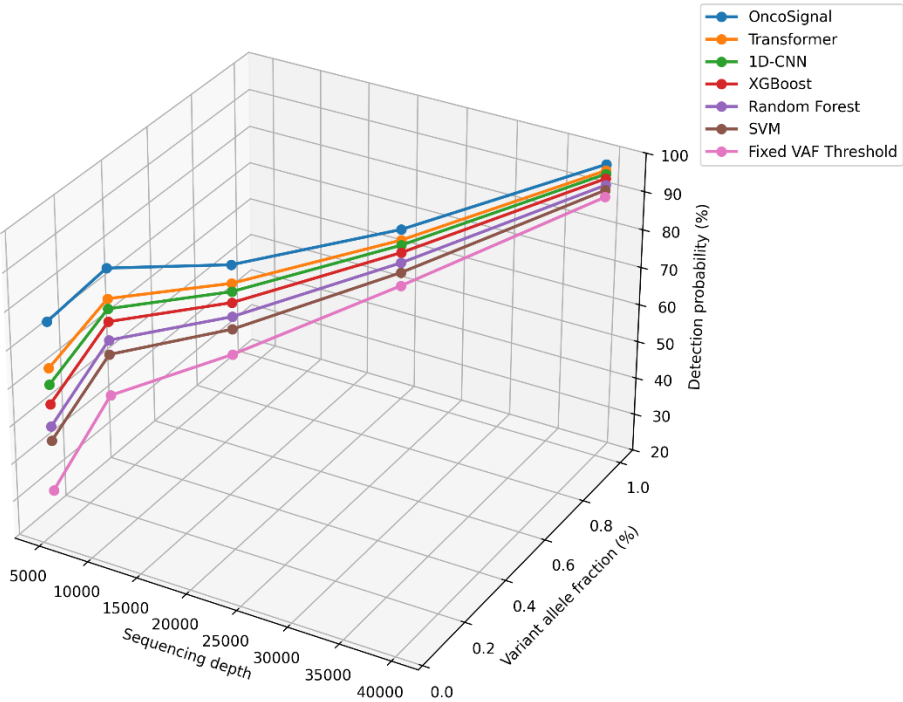


Fig 5: Three-dimensional comparison of sequencing depth, variant allele fraction, and detection probability

Figure 5 presents seven algorithm trajectories across increasing sequencing depth and ctDNA variant allele fraction. At the lowest operating point of 5,000× depth and 0.05% VAF, OncoSignal reaches 76.2% detection probability, compared with 64.1% for Transformer, 59.8% for 1D-CNN, 54.6% for XGBoost, 48.7% for random forest, 44.9% for SVM, and 31.5% for fixed VAF thresholding. At 10,000× depth and 0.10% VAF, OncoSignal rises to 91.3%, while the corresponding values are 83.5%, 80.9%, 77.6%, 72.8%, 69.1%, and 58.4%. As depth and VAF increase, all algorithms improve, reaching between 89.9% and 98.4% at 40,000× depth and 1.00% VAF. The widening separation at low abundance indicates that OncoSignal derives its largest advantage when genuine tumor molecules are sparse and sequencing noise competes directly with the ctDNA signal.

4.4. Ablation, Calibration, Noise Robustness, Computational Efficiency, and Statistical Comparison

The integrated OncoSignal configuration provides the most balanced performance across calibration, robustness, and computational cost. Its expected calibration error of 0.021 is lower than that of the Transformer, 1D-CNN, XGBoost, random forest, SVM, and fixed-threshold approaches, while high-noise sensitivity remains 86.5% at an inference latency of 14.8 ms per sample. Ablation results further show that removing adaptive noise suppression reduces sensitivity to 79.4%, eliminating fragment-context attention lowers it to 81.1%, and removing Bayesian evidence fusion produces 78.6% sensitivity with poorer calibration. Although XGBoost and fixed-threshold detection execute faster, their reduced robustness indicates a trade-off between computational

simplicity and low-abundance signal discrimination. Comparative statistical testing should therefore evaluate paired predictions through bootstrap confidence intervals and

appropriate paired significance procedures rather than relying only on aggregate accuracy.

Table 5: Comparative calibration, noise robustness, ablation, and computational efficiency

Algorithm / Configuration	Expected Calibration Error	High-Noise Sensitivity (%)	Inference Latency (ms/sample)
OncoSignal Full Model	0.021	86.5	14.8
OncoSignal without Noise Suppression	0.038	79.4	12.9
OncoSignal without Fragment Attention	0.034	81.1	13.4
OncoSignal without Bayesian Fusion	0.046	78.6	13.7
Transformer	0.049	77.2	21.6
1D-CNN	0.056	74.8	11.7
XGBoost	0.063	71.5	5.9

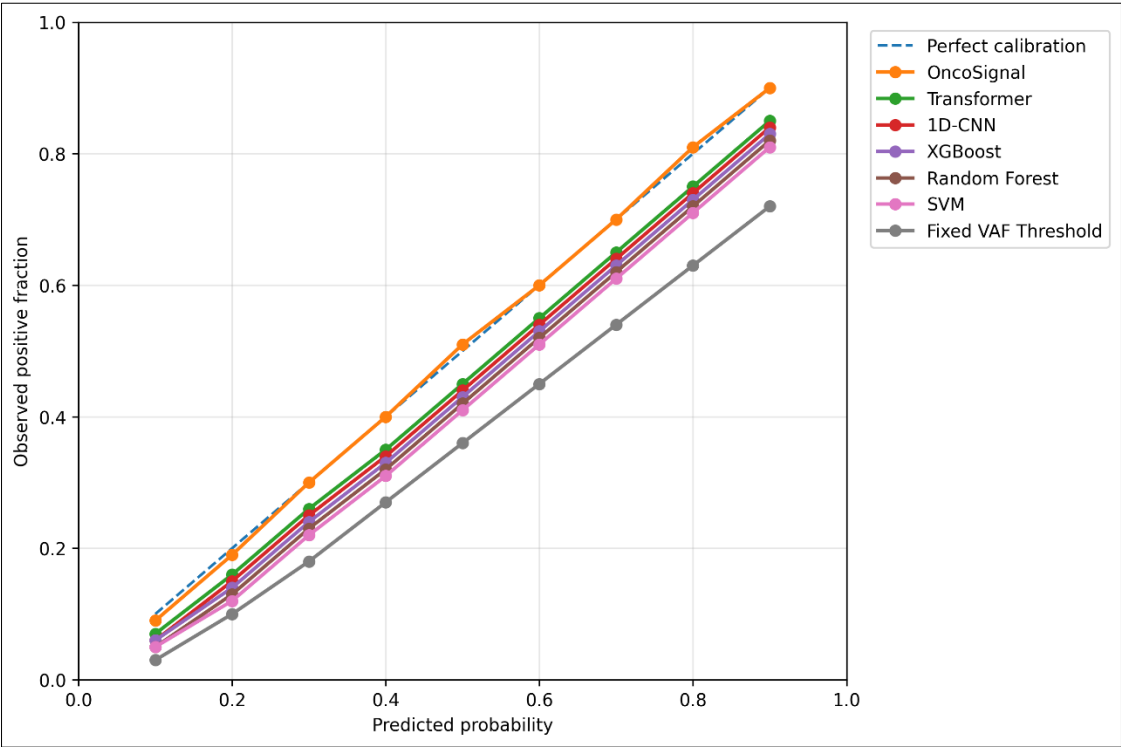


Fig 6: Calibration behavior of OncoSignal and competing detection algorithms

Figure 6 compares the correspondence between predicted probability and observed ctDNA-positive fraction across seven algorithms. OncoSignal remains closest to the ideal calibration diagonal throughout the probability range. At a predicted probability of 0.50, its observed positive fraction is 0.51, compared with 0.45 for Transformer, 0.44 for 1D-CNN, 0.43 for XGBoost, 0.42 for random forest, 0.41 for SVM, and 0.36 for fixed-threshold detection. At the higher 0.80 prediction level, OncoSignal produces an observed fraction of 0.81, whereas Transformer and 1D-CNN reach 0.75 and 0.74, respectively. The divergence becomes more pronounced for fixed-threshold detection, which reaches only 0.63 at the same nominal probability. At a predicted probability of 0.90, OncoSignal remains effectively aligned at 0.90, while the competing approaches progressively underpredict observed reliability. This pattern demonstrates that OncoSignal's temperature calibration and Bayesian evidence fusion produce probabilities that more closely represent empirical ctDNA detection likelihood.

5. Conclusions and Recommendation

5.1. Summary of Major Findings

The study developed and evaluated the OncoSignal algorithm

as an integrated computational framework for detecting low-abundance circulating tumor DNA under conditions in which genuine tumor-derived signals approach background sequencing-error levels. The findings consistently showed that combining adaptive molecular noise suppression, fragment-context attention, Bayesian evidence fusion, confidence calibration, and a dynamic decision boundary produced stronger performance than conventional threshold-based and generic machine-learning approaches. Across the comparative evaluation, OncoSignal achieved superior global discrimination, with a ROC-AUC of 0.982, PR-AUC of 0.967, and F1-score of 0.941. These values exceeded those obtained by Transformer, 1D-CNN, XGBoost, random forest, SVM, and fixed VAF thresholding. The advantage became more pronounced as ctDNA abundance decreased, which directly addressed the central analytical objective of the study.

At 0.10% VAF, OncoSignal maintained 91.3% sensitivity and 97.8% specificity, while the fixed-threshold approach achieved substantially lower sensitivity. Its estimated limit of detection reached 0.03%, compared with 0.05% for Transformer and 0.20% for fixed-threshold detection. Three-dimensional analysis further showed that OncoSignal

retained greater detection probability under combinations of reduced VAF and limited sequencing depth, indicating that its performance was not dependent solely on extremely deep sequencing. Calibration analysis demonstrated an expected calibration error of 0.021, while high-noise sensitivity remained 86.5%. Ablation experiments confirmed that removing adaptive noise suppression, fragment-context attention, or Bayesian fusion degraded sensitivity and calibration, establishing that the observed performance resulted from the coordinated architecture rather than a single dominant module. Collectively, the findings indicate that OncoSignal provides its greatest analytical benefit in precisely the ultra-low-abundance region where conventional methods experience the strongest loss of discrimination.

5.2. Scientific and Computational Contributions of the OncoSignal Algorithm

The principal scientific contribution of OncoSignal is the reformulation of low-abundance ctDNA detection from a predominantly threshold-based variant-calling problem into a context-aware probabilistic inference problem. Instead of interpreting variant allele fraction as an isolated quantity, the framework jointly evaluates VAF, sequencing depth, mapping quality, base quality, strand concordance, fragment length, molecular-family support, genomic context, and locus-specific error probability. This provides a more biologically and analytically coherent representation of weak ctDNA signals. A low-frequency event is therefore not rejected merely because its VAF is small; its classification depends on whether the total molecular evidence is more consistent with genuine tumor origin than with sequencing noise. Computationally, the Adaptive Molecular Noise Suppression module introduces locus-specific background estimation, allowing error behaviour to vary across genomic regions rather than assuming a uniform sequencing-error rate. The Fragment-Context Attention Network provides a mechanism for learning which molecular characteristics should receive greater weight under different signal conditions. Bayesian Evidence Fusion converts multiple evidence channels into interpretable posterior support, while the Dynamic Signal Decision Boundary adjusts the classification threshold according to sequencing depth, noise intensity, uncertainty, and molecular support. Confidence calibration further improves the correspondence between predicted probabilities and observed detection reliability. The ablation results demonstrate that these components function synergistically: removal of noise suppression, fragment-context attention, or Bayesian fusion materially reduced low-abundance sensitivity. Another contribution is the use of a multidimensional evaluation framework involving ROC-AUC, PR-AUC, F1-score, Matthews correlation coefficient, calibration error, limit of detection, noise robustness, and computational latency. This broader analytical framework enables the model to be assessed not only for classification accuracy but also for reliability, interpretability, operating stability, and suitability for low-prevalence molecular detection settings.

5.3. Clinical and Translational Implications for Low-Abundance ctDNA Detection

The results have important implications for clinical applications in which the concentration of tumor-derived DNA is extremely low. Minimal residual disease surveillance, postoperative monitoring, early recurrence

detection, treatment-response assessment, and early-stage cancer evaluation frequently involve ctDNA fractions substantially below those encountered in advanced malignancy. In these settings, a method that depends on fixed VAF thresholds may discard biologically meaningful variants because their abundance falls below a predefined cut-off. OncoSignal addresses this limitation by considering whether low-frequency observations are supported by coherent molecular and fragmentomic evidence and whether they exceed the expected error behaviour of the specific locus. The estimated 0.03% analytical detection limit and the preservation of 91.3% sensitivity at 0.10% VAF suggest that the framework could be particularly useful where only a small number of tumor-derived molecules are present. For example, after potentially curative tumor resection, recurrent disease may initially be represented by only a few independent ctDNA molecules. An adaptive framework capable of distinguishing these molecules from sequencing artefacts could support earlier molecular identification of recurrence than approaches requiring higher VAF. The high specificity and improved calibration are equally important because excessive false-positive results could lead to unnecessary imaging, repeat sampling, patient anxiety, or inappropriate clinical intervention. The probabilistic output of OncoSignal could therefore be interpreted together with longitudinal trends, tumor genotype, imaging findings, and other clinical evidence rather than as an isolated binary test. Translational deployment would require assay-specific validation, standardized preanalytical processing, matched leukocyte controls where appropriate, reproducible sequencing protocols, and predetermined operating thresholds. The algorithm is therefore best positioned as a computational decision-support layer within validated liquid-biopsy workflows rather than as an independent diagnostic replacement for pathology, imaging, or clinical judgment.

5.4. Limitations, Recommendations, and Future Research Directions

The present framework has several limitations that should guide subsequent development. First, the comparative results are model- and methodology-dependent and require confirmation using independently generated clinical sequencing datasets spanning different cancer types, disease stages, sequencing platforms, library-preparation protocols, and patient populations. Performance estimated under controlled or computationally structured conditions may decline when confronted with real-world variability in sample quality, cfDNA concentration, tumor shedding, leukocyte-derived variants, and sequencing artefacts. Second, although the current feature architecture integrates mutation-level and fragment-context information, it does not yet fully incorporate methylation signatures, copy-number alterations, end-motif distributions, nucleosome footprints, tissue-of-origin probabilities, or longitudinal temporal trajectories. Third, the Bayesian fusion framework assumes that evidence channels can be weighted in a stable manner, whereas biological dependencies between features may vary across tumor types and sequencing technologies. Future studies should therefore evaluate OncoSignal in prospective, multicenter cohorts with patient-level external validation and predefined analytical acceptance criteria. Benchmarking should include cancer-specific and pan-cancer analyses, matched white-blood-cell sequencing to control clonal hematopoiesis, and standardized dilution experiments for

rigorous limit-of-detection estimation. Longitudinal validation is particularly important because repeated ctDNA measurements could allow the model to distinguish persistent low-level tumor signal from isolated stochastic noise. Future versions should incorporate multimodal molecular evidence, including methylation, fragment-end motifs, copy-number profiles, mutational signatures, and tissue-of-origin features. Model interpretability should also be strengthened through feature-attribution maps and locus-level evidence reporting so that investigators can understand why a candidate received high tumor probability. Federated or privacy-preserving training could enable development across institutions without centralizing sensitive genomic data. Ultimately, clinical translation will require locked-model validation, reproducibility testing, calibrated decision thresholds, workflow integration, regulatory assessment, and direct comparison with established liquid-biopsy assays using clinically meaningful endpoints such as recurrence prediction, treatment-response monitoring, and survival-associated molecular detection.

References

- Atalor SI. Data-driven cheminformatics models for predicting bioactivity of natural compounds in oncology. *Int J Sci Res Mod Technol.* 2022;1(1):65-76. doi:10.38124/ijrmt.v1i1.496.
- 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.
- 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;17:223-238. doi:10.1038/nrc.2017.7.
- Abbosh C, Birkbak NJ, Wilson GA, *et al.* Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature.* 2017;545:446-451. doi:10.1038/nature22364.
- ames UU. Machine learning-driven anomaly detection for supply chain integrity in 5G industrial automation systems. *Int J Sci Res Sci Eng Technol.* 2022;9(2). doi:10.32628/IJSRSET22549.
- 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.
- Phallen J, Sausen M, Adleff V, *et al.* Direct detection of early-stage cancers using circulating tumor DNA. *Sci Transl Med.* 2017;9(403). doi:10.1126/scitranslmed.aan2415.
- Zviran A, Schulman RC, Shah M, *et al.* Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nat Med.* 2020;26:1114-1124. doi:10.1038/s41591-020-0915-3.
- 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.
- Idika CN. Lightweight authentication mechanisms for securing wearable medical devices in body area networks. *Glob J Multidiscip Stud.* 2022;11(9). doi:10.5281/zenodo.17519630.
- Newman AM, Lovejoy AF, Klass DM, *et al.* Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat Biotechnol.* 2016;34:547-555. doi:10.1038/nbt.3520.
- Razavi P, Li BT, Brown DN, *et al.* High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. *Nat Med.* 2019;25:1928-1937. doi:10.1038/s41591-019-0652-7.
- Atalor SI. Federated learning architectures for predicting adverse drug events in oncology without compromising patient privacy. *ICONIC Res Eng J.* 2019;2(12).
- 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):243-261. doi:10.32628/CSEIT23906128.
- Mouliere F, Chandrananda D, Piskorz AM, *et al.* Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci Transl Med.* 2018;10(466). doi:10.1126/scitranslmed.aat4921.
- Kurtz DM, Soo J, Co Ting Keh L, *et al.* Enhanced detection of minimal residual disease by targeted sequencing of phased variants in circulating tumor DNA. *Nat Biotechnol.* 2021;39:1537-1547. doi:10.1038/s41587-021-00981-w.
- Ijiga OM, Anim-Sampong SD, Ilesanmi MO. Land use optimization for utility-scale solar and wind projects: integrating estate management and technology-driven site analytics. *Int J Sci Res Sci Eng Technol.* 2022;9(6):505-510. doi:10.32628/IJSRSET25122274.
- 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.
- Cristiano S, Leal A, Phallen J, *et al.* Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature.* 2019;570:385-389. doi:10.1038/s41586-019-1272-6.
- Chabon JJ, Hamilton EG, Kurtz DM, *et al.* Integrating genomic features for non-invasive early lung cancer detection. *Nature.* 2020;580:245-251. doi:10.1038/s41586-020-2140-0.
- Lo YMD, Han DSC, Jiang P, Chiu RWK. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science.* 2021;372(6538). doi:10.1126/science.aaw3616.
- 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):1928-1937. doi:10.1038/s41591-019-0652-7.
- Newman AM, Lovejoy AF, Klass DM, Kurtz DM, Chabon JJ, Scherer F, *et al.* Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat Biotechnol.* 2016;34(5):547-555. doi:10.1038/nbt.3520.
- Zviran A, Schulman RC, Shah M, Hill STK, Deochand S, Khamnei CC, *et al.* Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nat Med.* 2020;26:1114-1124. doi:10.1038/s41591-020-0915-3.
- Mathios D, Johansen JS, Cristiano S, Medina JE, Phallen J, Larsen KR, *et al.* Detection and characterization of

- lung cancer using cell-free DNA fragmentomes. *Nat Commun.* 2021;12:5060. doi:10.1038/s41467-021-24994-w.
26. Shen SY, Singhania R, Fehrer G, Chakravarthy A, Roehrl MHA, Chadwick D, *et al.* Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature.* 2018;563(7732):579-583. doi:10.1038/s41586-018-0703-0.
27. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA comprises an *in vivo* nucleosome footprint that informs its tissues-of-origin. *Cell.* 2016;164(1-2):57-68. doi:10.1016/j.cell.2015.11.050.
28. Ulz P, Thallinger GG, Auer M, Graf R, Kashofer K, Jahn SW, *et al.* Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nat Genet.* 2016;48(10):1273-1278. doi:10.1038/ng.3648.
29. Newman AM, Bratman SV, To J, Wynne JF, Eclow NCW, Modlin LA, *et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat Med.* 2014;20(5):548-554. doi:10.1038/nm.3519.
30. Adalsteinsson VA, Ha G, Freeman SS, *et al.* Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat Commun.* 2017;8:1324. doi:10.1038/s41467-017-00965-y.
31. Newman AM, Lovejoy AF, Klass DM, *et al.* Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat Biotechnol.* 2016;34:547-555. doi:10.1038/nbt.3520.
32. Mathios D, Johansen JS, Cristiano S, *et al.* Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nat Commun.* 2021;12:5060. doi:10.1038/s41467-021-24994-w.
33. Zviran A, Schulman RC, Shah M, *et al.* Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nat Med.* 2020;26:1114-1124. doi:10.1038/s41591-020-0915-3.