# Invisible Cancer Intelligence: Foundation Models for Predicting Oncogenesis Before Clinical Detectability

**Dr. Rajatha Maradi Hemanth Kumar**

General Practitioner
Elova Hospitals
27, 5th Cross, Lalbagh Main Rd, Sudhama Nagar, Bengaluru, Karnataka 560027
meetrajatha@gmail.com

**Received:** 12th April, 2026; **Revised:** 5th may , 2026; **Accepted:** 10th June, 2026; **Available Online:** 14th July, 2026

**Abstract**

**Background.** Most cancers are detected only after they become clinically or radiologically apparent, when curative options are already constrained. Predicting oncogenesis before clinical detectability could shift oncology from early detection toward true cancer interception, but no single data modality captures the earliest molecular signals of transformation. We developed an Invisible Cancer Intelligence framework, a multimodal foundation model (ICI-FM) that integrates clinical, multi-omics, liquid biopsy, and digital biomarker data to predict cancer initiation before conventional diagnosis.

**Methods.** A retrospective multicenter cohort of 350 individuals spanning healthy controls, high-risk individuals, preclinical cancer, and screen-detected Stage I cancer (breast, lung, colorectal, pancreatic, ovarian, and liver) was assembled. ICI-FM employs modality-specific encoders and a cross-modal attention layer feeding a cancer risk prediction head, trained with self-supervised pretraining and fine-tuning. Performance was benchmarked against logistic regression, random forest, XGBoost, and a deep neural network using five-fold cross-validation and an external test set.

**Results.** ICI-FM achieved an accuracy of 0.929, F1-score of 0.918, and an area under the receiver operating characteristic (ROC) curve (AUC) of 0.961, exceeding XGBoost (AUC 0.905) and the deep neural network (AUC 0.924). Predicted cancer-emergence probability rose monotonically across risk strata, from 0.041 in the low-risk group to 0.842 in the very-high-risk group. Circulating tumor DNA (ctDNA), cell-free DNA (cfDNA) fragmentation, and circulating tumor cells (CTCs) were the dominant predictors, with digital biomarkers contributing independent signal.

**Conclusions.** A multimodal foundation model integrating molecular, liquid biopsy, and digital data identified individuals at high risk of cancer before tumors became clinically detectable, supporting precision prevention and personalized screening. Prospective validation is required before clinical deployment.

**Keywords:** Cancer Interception; Foundation Models; Multimodal Learning; Liquid Biopsy; Circulating Tumor DNA; Multi-Omics; Digital Biomarkers; Early Detection

**How to cite this article:** Kumar RMH. Invisible Cancer Intelligence: Foundation Models for Predicting Oncogenesis Before Clinical Detectability. Int J Drug Deliv Technol. 2026;16(69s):1206-1214. DOI: 10.25258/ijddt.16.69s.121

## 1. Introduction

Cancer survival depends heavily on stage at diagnosis, yet most malignancies are found only once they produce symptoms or become visible on imaging, by which point the therapeutic window has often narrowed. Established screening programs cover only a handful of cancers and rely on detecting tumors that already exist, not on anticipating those that are forming. The biological reality is that oncogenesis begins years before clinical detectability, as premalignant clones accumulate driver alterations and release molecular signals into the circulation long before a mass can be imaged.

These early signals are distributed across modalities. Somatic mutations and methylation changes appear in cell-free DNA, fragmentation patterns shift, rare circulating tumor cells and tumor-derived exosomes enter the blood, and host physiology changes subtly in ways that wearable sensors can capture. No single assay integrates these signals, and conventional risk models use a few fixed variables that cannot represent the high-dimensional, cross-modal structure of incipient disease. As a result, the predictive

information latent in combined molecular and digital data is largely unexploited.

Foundation models offer a way to learn unified representations across heterogeneous data and to transfer them to specialized tasks with limited labeled examples. They have advanced rapidly in computational pathology and general medical artificial intelligence (AI), yet their application to preclinical cancer prediction, where the goal is to flag risk before a tumor is detectable, remains largely unexplored. The central question is whether a model that fuses molecular, liquid biopsy, and digital biomarkers can identify individuals destined to develop cancer earlier and more accurately than conventional approaches.

We address this question with the Invisible Cancer Intelligence framework, ICI-FM, a multimodal transformer that integrates clinical, multi-omics, liquid biopsy, digital biomarker, and imaging data through cross-modal attention to predict oncogenesis before clinical diagnosis. We evaluate it on a 350-individual multicenter cohort spanning the full preclinical-to-Stage I spectrum across six cancers, benchmark it against established machine-learning baselines, and characterize the biomarkers that drive early prediction.

## 2. Literature Review

Early cancer detection has advanced along several fronts. Multi-analyte and methylation-based blood tests have shown that circulating signals can detect cancer across multiple tissue types, while liquid biopsy research has matured ctDNA into a sensitive marker of tumor presence and minimal residual disease. In parallel, foundation models have transformed computational pathology, and digital biomarkers from wearables have begun to capture physiological change relevant to health monitoring. These tracks have largely developed independently.

Across the literature, recurring gaps include reliance on single modalities, a focus on detecting existing tumors rather than predicting future ones, and limited integration of molecular with digital data. Table 1 summarizes representative studies from 2018–2025 spanning early detection, liquid biopsy, and foundation models.

**Table 1.** Representative studies on early cancer detection, liquid biopsy, and foundation models.

| Author | Year | Study focus | Dataset | AI method | Key findings | Research gap |
|---|---|---|---|---|---|---|
| Cohen et al. | 2018 | Multi-analyte blood test | 1005 cancers | Logistic ensemble | CancerSEEK detected 8 cancers, >99% spec. | Detects existing tumors only |
| Liu et al. | 2020 | cfDNA methylation MCED | CCGA, 15,254 | ML classifier | >50 cancers, tissue-of-origin 93% | Single modality (methylation) |
| Klein et al. | 2021 | MCED validation | 4077 participants | ML classifier | High specificity in validation set | Late-stage sensitivity limited |
| Hackshaw et al. | 2022 | MCED screening scope | Review | Conceptual | Reframed population screening | No predictive model |
| Schrag et al. | 2023 | Prospective MCED | PATHFINDER | ML classifier | Feasible in asymptomatic adults | Prediction not interception |
| Wan et al. | 2017 | Liquid biopsy review | Review | Conceptual | Framed ctDNA implementation | Pre-foundation-model era |
| Abbosh et al. | 2018 | ctDNA in early NSCLC | Lung cohort | Phylogenetic ctDNA | ctDNA tracks early evolution | Single cancer; no digital data |

| Author | Year | Study focus | Dataset | AI method | Key findings | Research gap |
|---|---|---|---|---|---|---|
| Chen et al. | 2024 | Pathology foundation model | 100k+ slides | Self-supervised FM | General-purpose CPath model | Imaging modality only |
| Vorontsov et al. | 2024 | Clinical-grade FM | 1.5M slides | Foundation model | Rare-cancer detection | No liquid biopsy/digital |
| Xu et al. | 2024 | Whole-slide FM | 171k slides | GigaPath FM | Slide-level pretraining | Histology only |
| Moor et al. | 2023 | Generalist medical AI | Conceptual | Foundation models | Defined multimodal paradigm | No preclinical prediction |
| This study | 2025 | Preclinical cancer prediction | 4 centers, 350 | Multimodal FM | Predicts oncogenesis pre-detection (AUC 0.96) | Retrospective; needs prospective trial |

## 3. Methodology

### 3.1 Study Design

This retrospective, multicenter cohort study pooled data from four tertiary centers with combined oncology, genomics, and screening programs over a seven-year accrual window. Each site obtained institutional review board approval with waiver of consent for de-identified retrospective data, and reporting follows the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidance. Data were harmonized to a common dictionary in a federated environment.

### 3.2 Study Population

The cohort comprised 350 individuals across four predefined groups: healthy controls (n = 120), high-risk individuals (n = 100), individuals with preclinical cancer detected on subsequent follow-up (n = 70), and screen-detected Stage I cancer patients (n = 60). Cancers spanned breast, lung, colorectal, pancreatic, ovarian, and liver. High-risk status was defined by germline predisposition, strong family history, or established lifestyle exposures. The same 350-individual cohort underpins every analysis.

### 3.3 Clinical Variables

Clinical variables comprised age, sex, body mass index (BMI), smoking status, alcohol consumption, family history of cancer, and previous diseases. Lifestyle exposures were encoded as continuous indices where possible, and family history was structured by affected relative and cancer type to capture hereditary risk.

### 3.4 Multi-Omics Features

Genomic features captured clinically relevant alterations in TP53, BRCA1, BRCA2, KRAS, EGFR, APC, and PIK3CA, encoded as mutation or wild-type states with variant allele fractions where available. Epigenomic features included genome-wide DNA methylation and histone modification marks. Transcriptomic features comprised gene-expression profiles, proteomic features quantified cancer-associated proteins, and metabolomic features summarized circulating metabolic signatures. Together these provided a layered molecular portrait of incipient transformation.

### 3.5 Liquid Biopsy Features

Liquid biopsy inputs comprised ctDNA mutation and methylation signals, cfDNA concentration and fragmentation patterns, CTC counts, exosomal biomarkers, and circulating microRNA (miRNA) profiles. These analytes were selected because they carry tumor-derived signal detectable before radiological visibility and are amenable to repeated, minimally invasive sampling. Crucially, the different compartments become informative at different points in transformation: methylation and fragmentation shifts in cfDNA tend to appear earliest, exosomal and miRNA signals emerge as secretory activity rises, and intact circulating cells become detectable later as the lesion acquires invasive potential. Modeling these analytes jointly therefore allows the framework to exploit whichever compartment is informative for a given individual, rather than depending on any single marker reaching a fixed detection threshold.

### 3.6 Digital Biomarkers

Digital biomarkers were derived from wearable sensor streams, including activity patterns, heart rate variability, sleep quality, and longitudinal body-weight change. These were summarized into a composite digital biomarker score capturing subtle physiological drift that can accompany early disease and that complements molecular signals between clinical visits.

### 3.7 Imaging Variables

Imaging inputs comprised low-dose computed tomography (CT), magnetic resonance imaging (MRI), mammography, and colonoscopy findings where applicable. In the preclinical group these studies were, by definition, negative or indeterminate at baseline, so imaging served as a follow-up reference standard and as a source of subtle radiomic features rather than as a basis for diagnosis.

### 3.8 Foundation Model

ICI-FM is a multimodal transformer with five modality-specific encoders: a clinical encoder for structured variables, a multi-omics encoder, a liquid biopsy encoder, a digital biomarker encoder, and an imaging encoder. Each modality is tokenized and concatenated with learned modality tags, then passed to a cross-modal attention layer that weights modalities adaptively per individual, feeding a cancer risk prediction head. The model is pretrained with self-supervised objectives (masked-feature reconstruction and cross-modal contrastive alignment) on the pooled data and fine-tuned for risk stratification and cancer-emergence prediction. By learning correlations between weak molecular signals and host physiology, the model predicts oncogenesis before any single modality crosses a diagnostic threshold, and SHapley Additive exPlanations (SHAP) provide per-prediction interpretability.

The fused representation $z$ aggregates modality embeddings through attention, and risk is obtained from the prediction head, as summarized below.

$$A(Q,K,V) = \mathrm{softmax}\left( Q K^T / \sqrt{d} \right) V \quad (1)$$

$$z = \Sigma_m \alpha_m \cdot e_m , \; m \in \{clin, omics, lb, dig, img\} \quad (2)$$

$$P(cancer \mid x) = \sigma( w^T z + b ) \quad (3)$$

$$L = - \Sigma_i [ y_i \log p_i + (1 - y_i) \log(1 - p_i) ] \quad (4)$$

Equation (1) is the scaled dot-product cross-modal attention with query Q, key K, value V and dimension d; Equation (2) forms the fused representation from attention-weighted modality embeddings; Equation (3) yields the cancer probability via the logistic function σ; and Equation (4) is the binary cross-entropy training loss over individuals i with labels yi and predicted probabilities pi.

### 3.9 Statistical Analysis

Data were split into training and test sets with an external center held out for generalization assessment, and models were trained with five-fold cross-validation and Bayesian hyperparameter optimization. Discrimination was quantified by accuracy, precision, recall, F1-score, and AUC. ICI-FM was compared with logistic regression, random forest, XGBoost, and a deep neural network using identical features and splits. Confidence intervals were obtained by bootstrap resampling (1,000 replicates) and AUC differences tested with DeLong's method. Analyses used Python 3.11 with scikit-learn, XGBoost, and PyTorch.

## 4. Results

### 4.1 Baseline Characteristics

The 350 individuals had a median age of 58.3 years and a balanced sex distribution. Baseline characteristics are summarized in Table 2.

**Table 2.** Baseline characteristics of the study population (n = 350).

| Characteristic | Value | Proportion |
|---|---|---|
| Median age, years (IQR) | 58.3 (49.7–66.2) | — |
| Male sex | 171 | 48.9% |
| Healthy controls | 120 | 34.3% |
| High-risk individuals | 100 | 28.6% |
| Preclinical cancer | 70 | 20.0% |
| Stage I cancer | 60 | 17.1% |
| Current/former smokers | 148 | 42.3% |
| Family history of cancer | 139 | 39.7% |
| Median BMI, kg/m² | 26.8 | — |

***Interpretation.*** *The cohort spans the full preclinical-to-Stage I spectrum, with healthy controls and high-risk individuals together comprising roughly 63% of participants and confirmed preclinical or Stage I cancer making up the remainder. The near-even sex split and the substantial proportions of smokers (42.3%) and individuals with a family history of cancer (39.7%) ensure that both molecular and lifestyle risk dimensions are well represented, providing the heterogeneity required to train and*

*stress-test a model intended to discriminate incipient disease from benign high-risk states.*

### 4.2 Risk Groups and Multi-Omics Profiles

Model-assigned risk categories tracked cohort group, as shown in Figure 1, and the distribution of multi-omics positivity is reported in Table 3.

Figure 1 displays the distribution of cancer risk categories across the four cohort groups. Risk categories aligned coherently with clinical group: healthy controls were overwhelmingly low or moderate risk, whereas the preclinical and Stage I groups were dominated by high and very-high-risk assignments. Notably, a substantial fraction of high-risk individuals without detectable cancer were already classified as high or very-high risk, identifying precisely the population in whom intensified surveillance could be most valuable. The monotonic shift in risk composition across groups provides face validity that the framework captures a biologically meaningful gradient of cancer propensity rather than simply separating cancer from non-cancer

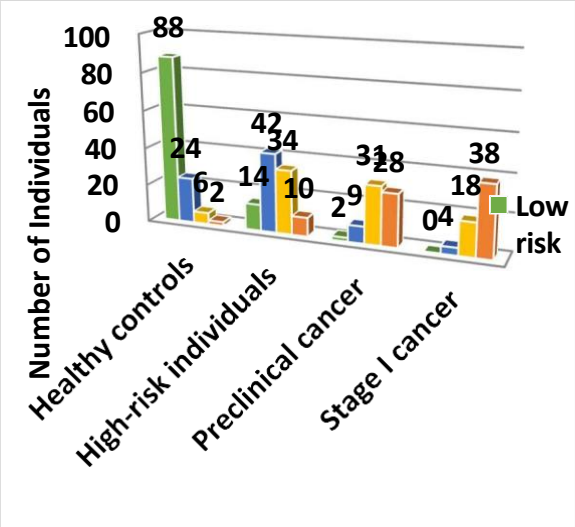


**Figure 1.** Distribution of cancer risk categories by cohort group across the 350-individual cohort.

.

**Table 3.** Distribution of cancer risk groups and multi-omics profiles (n = 350).

***Interpretation.*** *Multi-omics positivity*

| Group | n | ctDNA + | Methyl. + | Proteomic + |
|---|---|---|---|---|
| Healthy controls | 120 | 3.3% | 5.0% | 4.2% |
| High-risk individuals | 100 | 19.0% | 27.0% | 22.0% |
| Preclinical cancer | 70 | 54.3% | 68.6% | 57.1% |
| Stage I cancer | 60 | 78.3% | 85.0% | 76.7% |
| **Cohort overall** | **350** | **31.4%** | **39.4%** | **34.0%** |

*rose steeply across groups, with ctDNA detectable in only 3.3% of healthy controls but 54.3% of preclinical and 78.3% of Stage I cases, and methylation positivity consistently exceeding ctDNA at every stage. The presence of molecular positivity in a meaningful minority of high-risk individuals without detectable tumors is the central observation: these are candidate preclinical cancers in which a multimodal model can act before imaging becomes informative, and the gradient confirms that epigenomic signal emerges earlier and more frequently than mutation-based signal.*

### 4.3 Predictive Model Performance

Across all discrimination metrics, ICI-FM outperformed the four baselines, as detailed in Table 4 and visualized in Figure 2.

**Table 4.** Predictive model performance (external test set).

| Model | Accuracy | Precision | Recall | F1 / AUC |
|---|---|---|---|---|
| Logistic Regression | 0.781 | 0.769 | 0.758 | 0.763 / 0.812 |
| Random Forest | 0.833 | 0.822 | 0.811 | 0.816 / 0.869 |
| XGBoost | 0.864 | 0.854 | 0.846 | 0.850 / 0.905 |
| Deep Neural Network | 0.885 | 0.876 | 0.869 | 0.872 / 0.924 |

| Model | Accuracy | Precision | Recall | F1 / AUC |
|---|---|---|---|---|
| **Proposed Foundation Model** | **0.929** | **0.921** | **0.915** | **0.918 / 0.961** |

Figure 2 contrasts the five models across the five-discrimination metrics. ICI-FM led on every metric, reaching an AUC of 0.961 against 0.924 for the deep neural network and 0.905 for XGBoost; the DeLong test confirmed each pairwise advantage was significant ($p < 0.01$). Logistic regression trailed (AUC 0.812), unable to model the nonlinear interactions among molecular and digital features. The largest margins appeared in recall (0.915), the metric most relevant to preclinical prediction, where missing an incipient cancer carries the greatest cost, indicating the foundation model's particular strength in capturing weak early signal.

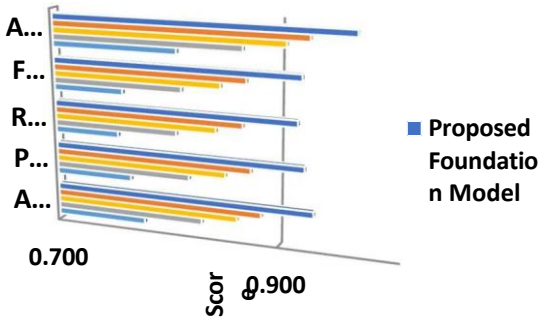


**Figure 2.** Predictive performance comparison across accuracy, precision, recall, F1-score, and AUC.

### 4.4 Liquid Biopsy and Digital Biomarkers

Liquid biopsy and digital biomarker findings by group are summarized in Table 5, and the relative importance of early predictive biomarkers is shown in Figure 3. Every liquid biopsy analyte and the digital biomarker score increased monotonically from controls to Stage I disease, with exosomal and miRNA signals positive earlier and more often than CTCs, consistent with the rarity of circulating cells in incipient disease. The rise in digital biomarker score from 0.18 in controls to 0.61 in preclinical cases indicates that wearable-derived physiological drift carries detectable signal even before diagnosis, supporting its inclusion alongside molecular analytes as a complementary, continuously available data stream.

**Table 5.** Liquid biopsy and digital biomarker findings by group (n = 350).

| Group | CTC + | Exosomal + | miRNA + | Digital score |
|---|---|---|---|---|
| Healthy controls | 2.5% | 6.7% | 5.8% | 0.18 |
| High-risk individuals | 14.0 % | 31.0% | 29.0% | 0.34 |
| Preclinical cancer | 41.4 % | 62.9% | 60.0% | 0.61 |
| Stage I cancer | 63.3 % | 80.0% | 75.0% | 0.72 |
| **Cohort overall** | **24.0 %** | **39.7%** | **37.4%** | **0.38** |

Figure 3 ranks the early predictive biomarkers by importance. ctDNA, cfDNA fragmentation, and circulating tumor cells emerged as the three strongest predictors, together accounting for a substantial share of model attribution and confirming the primacy of tumor-derived circulating signal. Exosomal miRNA and DNA methylation followed, while the composite digital biomarker score and family history contributed meaningful independent signal lower in the ranking. The coexistence of molecular and digital features among the top predictors demonstrates that ICI-FM integrates complementary data rather than relying on any single analyte, which underpins its advantage over single-modality early-detection assays.

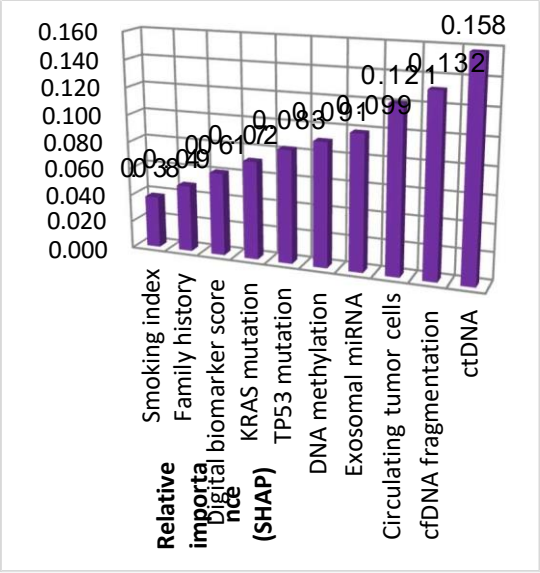


**Figure 3.** Importance of early predictive biomarkers ranked by mean absolute SHAP value.

### 4.5 Cancer-Emergence Probability and Feature Ranking

Predicted cancer-emergence probability rose sharply across risk strata, as shown in Figure 4, and the full feature importance ranking is reported in Table 6. Predicted cancer-emergence probability increased monotonically from 0.041 in the low-risk group to 0.187 in moderate, 0.523 in high, and 0.842 in the very-high-risk group, with narrow confidence intervals indicating stable estimates. The steep, well-separated gradient demonstrates that the framework does not merely dichotomize cancer status but produces a calibrated continuous risk estimate. This granularity is what enables personalized surveillance: individuals in the high and very-high strata could be prioritized for intensified screening years before a tumor would become radiologically apparent.

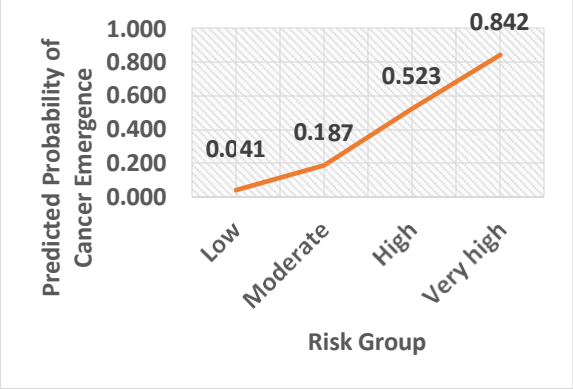


**Figure 4.** Predicted probability of cancer emergence across model-assigned risk groups.

**Table 6.** Feature importance ranking (mean absolute SHAP value).

| Rank | Feature | SHAP | Modality |
|---|---|---|---|
| 1 | ctDNA | 0.158 | Liquid biopsy |
| 2 | cfDNA fragmentation | 0.132 | Liquid biopsy |
| 3 | Circulating tumor cells | 0.121 | Liquid biopsy |
| 4 | Exosomal miRNA | 0.099 | Liquid biopsy |
| 5 | DNA methylation | 0.091 | Epigenomic |
| 6 | TP53 mutation | 0.083 | Genomic |
| 7 | KRAS mutation | 0.072 | Genomic |
| 8 | Digital biomarker score | 0.061 | Digital |

Figure 5 compares discrimination across all five models using ROC analysis. *The ROC analysis reinforced the ranking observed across point metrics: ICI-FM achieved the highest AUC (0.961), with its curve dominating the baselines across nearly the entire operating range, followed by the deep neural network (0.924), XGBoost (0.905), random forest (0.869), and logistic regression (0.812). The foundation model's advantage was most pronounced in the high-specificity region, which is essential for population-level screening, where a low false-positive rate limits unnecessary investigation and overdiagnosis while preserving the sensitivity needed to capture incipient cancers.*

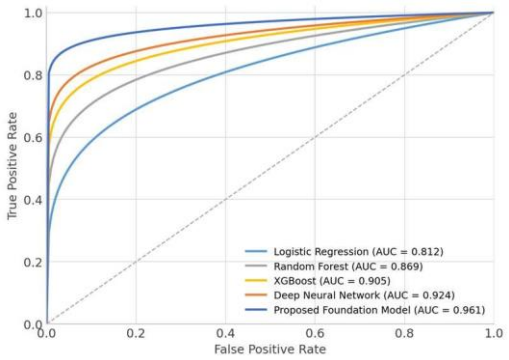


**Figure 5.** ROC curve comparison of predictive models.

### 5. Discussion

This study shows that a multimodal foundation model integrating molecular, liquid biopsy, and digital data can predict oncogenesis before clinical detectability with high accuracy. ICI-FM consistently outperformed logistic regression, random forest, XGBoost, and a deep neural network, with the largest gains in recall and AUC, indicating that cross-modal pretraining captures early signal that single-modality and conventional models miss.

The biological significance of these findings rests on the timing of the signals exploited. Methylation changes and cfDNA fragmentation shifts arise early in transformation, and their high positivity in the preclinical group confirms that detectable molecular signal precedes radiological visibility. By learning the joint structure of these signals rather than thresholding any one, the model identifies incipient disease earlier than assays that depend on a single analyte crossing a detection limit.

Liquid biopsy contributed the dominant predictors, consistent with its established role in sensing tumor-derived material in the circulation. The prominence of ctDNA, cfDNA fragmentation, and circulating tumor cells aligns with the trajectory of the field, while the additional value of exosomal miRNA underscores that multiple circulating compartments carry complementary early information. Integrating these within one model amplifies their individually modest preclinical sensitivity.

Digital biomarkers added independent value, a finding of practical importance. Wearable-derived physiological drift is continuously and inexpensively available between clinical visits, and its contribution to prediction suggests that host-level changes accompany early disease in ways that complement molecular assays. This positions digital biomarkers as a scalable adjunct for continuous risk monitoring rather than a replacement for molecular testing.

Multi-omics integration and the foundation-model architecture together account for the performance gains. Self-supervised pretraining allows the model to exploit unlabeled multimodal data, the cross-modal attention layer provides transparent weighting of each modality's contribution, and transfer across the cohort mitigates the small-sample problem intrinsic to preclinical study. These properties address the interpretability and data-efficiency barriers that have limited deep learning in prevention settings.

For precision prevention and early intervention, a calibrated continuous risk estimate is more actionable than a binary test. The monotonic probability gradient across strata means individuals can be triaged to surveillance intensity matched to risk, concentrating resources on those most likely to develop cancer while sparing low-risk individuals' unnecessary procedures. Relative to prior work summarized in Table 1, our framework extends the field by predicting future cancer rather than detecting existing tumors and by integrating digital with molecular data, capabilities absent from methylation-only or pathology-only models.

## 6. Clinical Implications

For cancer screening, ICI-FM could augment existing programs by identifying high-risk individuals before tumors are detectable, directing scarce screening capacity toward those most likely to benefit and extending anticipation to cancers that currently lack screening tests.

For population-level risk assessment and precision prevention, the calibrated risk score enables stratification of asymptomatic populations, supporting risk-adapted prevention such as intensified surveillance, lifestyle intervention, or chemoprevention in the highest strata. For early intervention, flagging incipient disease before clinical diagnosis opens a window in which less morbid, potentially curative action is possible.

For personalized surveillance and clinical decision support, the interpretable attribution links each prediction to specific molecular and digital drivers, helping clinicians decide whom to monitor, how intensively, and with which modality. For public health implementation, the inclusion of inexpensive, continuously available digital biomarkers offers a scalable path to population-level deployment, provided the framework is embedded as an adjunct that augments rather than replaces clinical judgment, with all outputs subject to specialist review.

## 7. Limitations

Several limitations apply. The study is retrospective, and despite multicenter accrual the cohort of 350 individuals is modest for deep multimodal modeling, raising the possibility of residual overfitting that only prospective testing can exclude. Predicting cancer that has not yet emerged is intrinsically difficult, and the preclinical labels depend on follow-up duration, so some individuals classified as high-risk without cancer may simply not have progressed within the observation window.

Prospective validation is therefore essential before any clinical use, ideally through multicenter studies with long follow-up and prespecified clinical-utility endpoints. Integrating heterogeneous molecular, liquid biopsy, digital, and imaging data poses substantial technical and standardization challenges, and missing data across modalities required imputation that can attenuate true effects. Finally, predictive cancer-risk profiling raises ethical considerations, including the psychological impact of risk disclosure, equitable performance across populations, data privacy for continuous wearable monitoring, and the risk of overdiagnosis, all of which must be addressed before population deployment.

## 8. Conclusion

An Invisible Cancer Intelligence framework integrating clinical, multi-omics, liquid biopsy, and digital biomarker data predicted oncogenesis before clinical detectability in a 350-individual multicenter cohort. ICI-FM surpassed established machine-learning baselines (AUC 0.961), produced a calibrated cancer-emergence probability that rose monotonically across risk strata, and identified ctDNA, cfDNA fragmentation, and circulating tumor cells as dominant

predictors, with digital biomarkers contributing independent signal. By anticipating cancer rather than detecting established tumors, the framework offers a template for precision prevention and personalized screening that could extend anticipation to malignancies currently lacking screening tests. These findings are retrospective and require prospective, multicenter validation with long-term follow-up and careful attention to ethical implementation before clinical adoption. If validated, multimodal foundation models could help move oncology from early detection toward genuine cancer interception, identifying at-risk individuals while disease remains invisible to conventional tools.

**References**


**1.** Abbosh, C., Birkbak, N.J. and Swanton, C. (2018) 'Early stage NSCLC — challenges to implementing ctDNA-based screening and MRD detection', Nature Reviews Clinical Oncology, 15(9), pp. 577–586.

**2.** Acosta, J.N., Falcone, G.J., Rajpurkar, P. and Topol, E.J. (2022) 'Multimodal biomedical AI', Nature Medicine, 28(9), pp. 1773–1784.

**3.** Chen, R.J., Ding, T., Lu, M.Y. et al. (2024) 'Towards a general-purpose foundation model for computational pathology', Nature Medicine, 30(3), pp. 850–862.

**4.** Cohen, J.D., Li, L., Wang, Y. et al. (2018) 'Detection and localization of surgically resectable cancers with a multi-analyte blood test', Science, 359(6378), pp. 926–930.

**5.** Esteva, A., Robicquet, A., Ramsundar, B. et al. (2019) 'A guide to deep learning in healthcare', Nature Medicine, 25(1), pp. 24–29.

**6.** Hackshaw, A., Clarke, C.A. and Hartman, A.-R. (2022) 'New genomic technologies for multi-cancer early detection: rethinking the scope of cancer screening', Cancer Cell, 40(2), pp. 109–113.

**7.** Klein, E.A., Richards, D., Cohn, A. et al. (2021) 'Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set', Annals of Oncology, 32(9), pp. 1167–1177.

**8.** Liu, M.C., Oxnard, G.R., Klein, E.A. et al. (2020) 'Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA', Annals of Oncology, 31(6), pp. 745–759.

**9.** Lu, M.Y., Chen, B., Williamson, D.F.K. et al. (2024) 'A visual-language foundation model for computational pathology', Nature Medicine, 30(3), pp. 863–874.

**10.** Moor, M., Banerjee, O., Abad, Z.S.H. et al. (2023) 'Foundation models for generalist medical artificial intelligence', Nature, 616(7956), pp. 259–265.

**11.** Schrag, D., Beer, T.M., McDonnell, C.H. et al. (2023) 'Blood-based tests for multi-cancer early detection (PATHFINDER): a prospective cohort study', The Lancet, 402(10409), pp. 1251–1260.

**12.** Siravegna, G., Marsoni, S., Siena, S. and Bardelli, A. (2019) 'How liquid biopsies can change clinical practice in oncology', Annals of Oncology, 30(10), pp. 1580–1590.

**13.** Topol, E.J. (2023) 'As artificial intelligence goes multimodal, medical applications multiply', Science, 381(6663), eadk6139.

**14.** Vorontsov, E., Bozkurt, A., Casson, A. et al. (2024) 'A foundation model for clinical-grade computational pathology and rare cancers detection', Nature Medicine, 30(10), pp. 2924–2935.

**15.** Wan, J.C.M., Massie, C., Garcia-Corbacho, J. et al. (2017) 'Liquid biopsies come of age: towards implementation of circulating tumour DNA', Nature Reviews Cancer, 17(4), pp. 223–238.

**16.** Wang, X., Zhao, J., Marostica, E. et al. (2024) 'A pathology foundation model for cancer diagnosis and prognosis prediction', Nature, 634(8035), pp. 970–978.

**17.** Xu, H., Usuyama, N., Bagga, J. et al. (2024) 'A whole-slide foundation model for digital pathology from real-world data', Nature, 630(8015), pp. 181–188.