The American Journal of Medical Sciences and Pharmaceutical Research ISSN 2689-1026 Volume 08 - 2026

A Multi-Omics Transformer Foundation Model For Ai-Driven Early Cancer Detection Using cfDNA And cfRNA: Implications For Precision Oncology And Early Intervention In U.S. Healthcare Systems

[1] Rabi Sankar Mondal*, [2] Reshad Aldin Ahmed, [3] Md Abu Kawsar Prodhan Hemal, [4] Tawfiqur Rahman Sikder, [5] Bismi Jatil Alia Juie

[1] Pompea College of Business, University of New Haven, West Haven, Connecticut, USA
[2] Department of Exercise Physiology, Central Michigan University, MI 48859, USA
[3] College of Computer Science, Pacific States University, Los Angeles, CA 90010, USA
[4] School of Business, International American University, Los Angeles, California, USA
[5] Training Department, Incepta Pharmaceuticals Limited, Dhaka-1208, Bangladesh

Received: 24th Dec 2025 | Received Revised Version: 25th Jan 2026 | Accepted: 18th Feb 2026 | Published: 21nd Feb 2026

Volume 08 Issue 02 2026 | Crossref DOI: 10.37547/tajmspr/Volume08Issue02-17

Abstract

A non-invasive and scalable substitute for tissue biopsies is liquid biopsy-based cancer diagnosis using circulating tumor DNA (cfDNA) and RNA (cfRNA); however, appropriate integration of heterogeneous molecular signals is still a major obstacle. Using TCGA gene expression and copy number variation data, we present a Transformer-based multi-omics fusion framework for pan-cancer identification in this paper. The suggested method dynamically weights informative molecular properties and learns intricate cross-omics interactions using self-attention. A single representation was produced by preprocessing, normalizing, and combining the gene expression and copy number variation profiles of 5,408 tumor samples from 33 different cancer types. The Autoencoder, Multilayer Perceptron, Support Vector Machine, and Logistic Regression baselines were used to compare the model. The Transformer achieved an accuracy of 82.6%, an F1-score of 82.0%, and an AUC of 0.899, surpassing all other methods. The results of this study demonstrate that attention-based multi-omics integration is a powerful and reliable method for liquid biopsy-based cancer diagnosis, which encourages the creation of scalable, AI-powered precision oncology systems.

Keywords: Liquid Biopsy; Multi-Omics Integration; Transformer-Based Deep Learning; Circulating Cell-Free DNA (cfDNA); Precision Oncology.

Cite This Article: Rabi Sankar Mondal, Reshad Aldin Ahmed, Md Abu Kawsar Prodhan Hemal, Tawfiqur Rahman Sikder, & Bismi Jatil Alia Juie. (2026). A Multi-Omics Transformer Foundation Model For Ai-Driven Early Cancer Detection Using cfDNA And cfRNA: Implications For Precision Oncology And Early Intervention In U.S. Healthcare Systems. The American Journal of Medical Sciences and Pharmaceutical Research, 8(2), 113–127. https://doi.org/10.37547/tajmspr/Volume08Issue02-17

1. Introduction

Cancer constitutes a major worldwide health issue because patient survival rates depend on the moment when doctors first identify their condition. Traditional diagnostic techniques based on imaging and biopsy procedures provide essential information for cancer diagnosis but these methods result in high costs and require invasive procedures while failing to identify early-stage cancers (Abreu et al., 2026). The non-invasive liquid biopsy technique which requires genetic biomarker analysis from peripheral blood samples has emerged as a solution to these problems

because it enables cancer detection during its initial development stage (Duan et al., 2026). The blood circulation contains circulating cell-free DNA (cfDNA) which cancer cells release through apoptosis and necrosis along with other cells. The detection of cfDNA occurs during early cancer stages because blood tests reveal low cancer cell percentages and scientists identify changes in fragment size and methylation status. Cell-free RNA (cfRNA) serves as a molecular converter that transmits real-time transcriptional data to display gene expression changes which occur during the early development stages of cancer. Research has shown that cfRNA profiling achieves high cancer detection accuracy when combined with next-generation sequencing and AI-based analysis methods (Zhong et al., 2024).

Scientists encounter challenges when trying to use cfDNA and cfRNA data because biological variability and random effects and different data structures create obstacles to their work. The high-dimensional data from multi-omics studies introduces greater difficulties because the system exhibits intricate relationships which traditional statistical methods and machine learning algorithms struggle to understand. The deep learning field now relies on transformer-based models as the main solution for solving complex prediction problems which require multiple data types. Researchers increasingly adopt deep learning techniques to support their work with multi-omics data by enabling data integration and feature discovery and prediction modeling (Sartori et al., 2025).

The self-attention mechanisms used in transformers make it possible for researchers to investigate how various data types interact with each other both inside their own categories and across different categories. The results of transformer-based systems show good performance when they combine multiple biological data sources which indicates that these systems can enhance diagnostic accuracy when they use cfDNA and cfRNA data (Pushparaj et al., 2026).

The study shows how cfDNA and cfRNA signals combine with deep learning technology to create cancer diagnostic tests which offer better detection performance and improved testing accuracy than current single-modality testing methods. The study introduces a Multi-Omics Transformer Foundation Model which uses artificial intelligence to identify early-stage cancer through simultaneous analysis of cfDNA and cfRNA data. Our research shows how U.S. healthcare systems can improve precision oncology practices through non-invasive screening methods which help identify patients needing early treatment.

2. Literature Review

Advances in liquid biopsy methods, fueled by progress in sequencing as well as AI, have significantly reshaped the landscape of early cancer detection. Liquid biopsy, with cell-free DNA (cfDNA), has evinced remarkable potential as an effective tool for non-invasive screening, enabling scientists to monitor signs of cancerous development by analyzing cellular mutations from a simple blood test. Genetic mutations, copy number variations (CNVs), methylation, fragmentation, as well as end motifs, are some of the key aspects of cfDNA with potential as supplementary data. Some scientific studies ascertained that fusion of different modes of cfDNA features enhances sensitivity as well as specificity in early detection compared to solely applying individual methods (Luo et al., 2025).

The development of cfRNA as an additional data source has created new research opportunities because it tracks real-time changes in gene activity which tumors undergo as they develop. The process of cfRNA profiling enables scientists to discover transcriptome signals which standard DNA testing would not find, thus giving them complete insights into disease development. The scientific field has developed better ways to analyze cfRNA, which exists in blood at lower levels, and this breakthrough enhances its use for identifying diseases in their initial stages because combined cfDNA-cfRNA testing produces better results than using each method separately (Zhong et al., 2024; Tanvir et al., 2024). Researchers from the past applied standard machine learning techniques which include logistic regression and support vector machines and ensemble techniques to classify cfDNA features. The study showed that multi-omics models successfully distinguished between cancer and non-cancer cases through their analysis of cfDNA-based characteristics according to the results from Abreu et al. (2026). The research showed that multi-omics machine learning models which combined cfDNA concentrations with CNVs and tumor markers produced better results than models which used only one type of feature according to the study results from Kwon et al., (2023).

However, recent developments in deep learning techniques such as transformers have shown greater potential in processing large dimensional and varying biological input data. In fact, transformer structures, originally developed for natural language processing, allow complex long-range dependencies and interactions between varying data modalities. In similar contexts, transformers have shown better results for multi-omics integration for other predictive

models of biomedical data, such as the integration of cfDNA and cfRNA for prediction of non-cancer conditions such as pre-term birth risk prediction, although superior prediction accuracy was obtained for multi-modal models than any single model, emphatically demonstrating synergies between diverse data integration. Despite their relative lack of adoption for any potential role in transformer-based causality prediction for cancer diagnosis, analogous examples from other domains predict the definitive role they will play in any multi-omics prediction of cancer (Zhou et al., 2025; Juie et al., 2021).

The performance of traditional machine learning algorithms which include support vector machines and random forests has shown only moderate success in multi-omics cancer classification. The models face difficulties because they cannot handle intricate connections present in diverse high-dimensional datasets (Song et al., 2025). The popularity of deep learning frameworks has increased because transformer topologies enable researchers to model extended time dependencies while working with various biological datasets. Transformer-based models demonstrate superior performance compared to traditional methods when analyzing cfDNA and cfRNA sequence data because they identify intricate molecular relationships. Integrated liquid biopsy techniques show their ability to translate research findings into clinical practice through evidence from extensive clinical studies.

The research from prospective efforts tested multi-omics classifiers which combined DNA methylation with mutation patterns and protein markers across multiple cancer types to prove that multimodal detection methods function effectively in real-world human populations. The combination of multi-omics features improves tissue-of-origin determination which serves as a critical component in multi-cancer early detection systems (Duan et al., 2026).

AI-based cancer research studies with greater sample sizes demonstrate the research benefits of using multiple omics data sources. The study demonstrates that multi-omics frameworks which combine genomic data with transcriptomic data improve cancer subtype identification and drug resistance prediction because integrated data fusion is essential for precise cancer treatment (Ran et al., 2025; Mohib et al., 2025). The study demonstrates that AI models use multiple omics data sources to create better biological understanding and medical applications even though it does not focus on cfDNA and cfRNA.

3. Methodology

This paper presents a multi-omics AI framework for early cancer detection that analyzes both circulating cell-free DNA (cfDNA) and circulating cell-free (cfDNA). The framework combines copy number variation (CNV) profiles from cfDNA and gene expression (GE) profiles from cfRNA into a single learning pipeline. The TCGA dataset provides cfDNA and cfRNA profiles which undergo normalization before feature selection processing begins. The two omics modalities are then encoded independently and combined using a Transformer model with cross-attention which enables the network to learn complex relations between genetic and transcriptomic data. The integrated representation is then utilized to construct cancer likelihood scores and classify samples as normal or cancerous, allowing for precise pan-cancer diagnosis. The cancer detection framework operates through a series of steps which Figure 1 displays as an overall process.

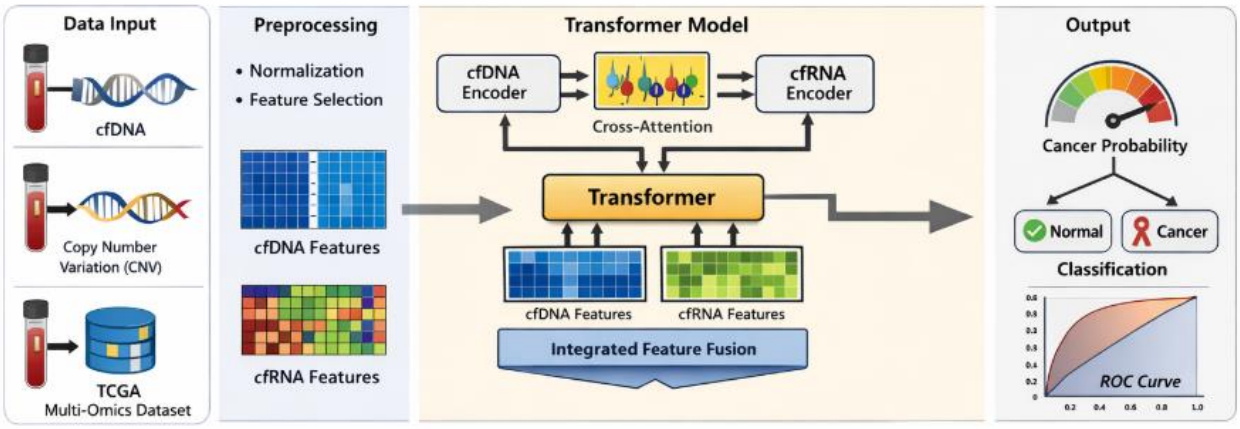

Figure 1. Proposed Transformer-Based Multi-Omics Fusion Architecture

3.1 Data acquisition

The research study employs a multi-omics pan-cancer dataset from "The Cancer Genome Atlas (TCGA)" which contains data about 5,408 tumor samples from 33 cancer types (Reymundez, 2020). The samples contain three molecular profiles which include gene expression (GE) and DNA methylation (METH) and copy number variation (CNV) that total 60,112 characteristics. The researchers obtained gene expression data through Illumina HiSeq RNA-Seq which they reported as log-transformed gene counts. The researchers used Illumina HM450 platform to produce DNA methylation data which they summarized at the CpG island level and converted to M-values. The CNV statistics show the intensity of copy number for each gene. The researchers conducted batch and tissue correction for all omics layers to achieve high-quality data and consistent biological results. The multi-omics profiles provide researchers with a reliable base which they can use to develop and evaluate cancer detection systems.

3.2 Data Preprocessing

Before the training of the model, the data was carefully preprocessed to ensure numerical stability. Additionally, it was preprocessed to reduce noise and increase the ability to effectively learn complex information. To begin with, the matrices containing Gene Expression (GE) data and Copy Number Variation (CNV) data were downloaded from the TCGA dataset. In addition, to ensure that each sample was aligned across different modalities such that each row represented the same sample across the cfDNA and cfRNA data, the data was aligned. To ensure that the data was not affected by differences in the magnitude or distribution of different molecular attributes, data normalization was performed. The data normalization of a given feature x is completed through the following steps:

$$x' = \frac{x - \mu}{\sigma}$$

The training set's mean and standard deviation are represented by μ and σ, respectively. This process helps ensure that each of the characteristics has an equivalent contribution in the learning of the models and that no information is dominated by other information. Since the original data set had tens of thousands of molecular characteristics, a feature selection approach based on the variance was utilized. This approach reduced the dimensions and eliminated some of the information while ensuring that the most important features, which represented the most physiological information, of the data set were selected. These would be the top variances of each of the omics layers. Once the preprocessing of the data was complete, i.e., the normalization and feature selection, a fused representation of the data set was achieved for each patient by uniting the two feature matrices of the cfDNA and cfRNA. The feature vectors generated from this uniting process were randomly split into the training set and the test set, each comprising 70% and 30%, respectively.

3.3 Transformer based Multi-Omics Fusion Model

The two molecular data types cfDNA copy number variations and cfRNA gene expression were combined to create a single, cohesive profile for every patient following preprocessing and feature selection.

Let,

$$\mathrm{X}^{DNA} \in \mathbb{R}^{n \times d_1}$$

denote the normalized CNV feature matrix and

$$\mathrm{X}^{RNA} \in \mathbb{R}^{n \times d_2}$$

denote the normalized gene expression matrix, where n is the number of samples and d_1,d_(2)are the selected feature dimensions.

These two matrices were concatenated at the feature level for the baselines of neural networks and traditional machine learning:

$$X = [\mathrm{X}^{DNA} \,||\, \mathrm{X}^{RNA}]$$

This early fusion process results in a single vector per sample, which combines both genomic instability (cfDNA) and transcriptional activity (cfRNA) information. This allows baseline models like LR, SVM, MLP, Autoencoder, etc., to be able to learn from these chemical sources in an early fusion manner.

The Transformer-based concept used a more complex fusion technique through its advanced fusion method. The research used separate latent embeddings to project cfDNA and cfRNA through modality-specific neuronal layers instead of directly concatenating their features.:

$$Z_{DNA} = f_{DNA}(\mathrm{X}^{DNA}),\ Z_{RNA} = f_{RNA}(\mathrm{X}^{RNA})$$

The two tokens which contained the processed embeddings were sent to a Transformer encoder for processing. The Transformer uses self-attention to learn the interaction of cfDNA and cfRNA characteristics which enables the model

to detect cross-omics connections that show how copy number variations impact gene expression. The Transformer creates a combined representation which contains information from both omics sources and this representation is used to classify cancer types. The attention-based fusion method enables the model to adjust its focus on each modality based on its value for cancer prediction which creates a more informative and biologically relevant integration method than simple concatenation.

The Transformer encoder produces a fused representation:

$$Z = Transformer([Z_{DNA}, Z_{RNA}])$$

This fused embedding captures both intra-omic and inter-omic relationships. The output is then passed through a fully connected classification head followed by a sigmoid activation function to estimate the probability of cancer:

$$\hat{y} = \sigma\,(WZ + b)$$

Unlike standard feature concatenation, Transformer-based design allows for variable weighting of omics modalities via attention. This enables the model to automatically evaluate whether cfDNA or cfRNA is more informative for a specific sample, hence increasing robustness and interpretability.

The proposed design, which explicitly models cross-omics interactions, is especially well-suited for liquid biopsy analysis, since cancer-related signals are weak and scattered over numerous molecular levels.

3.4 Training Strategy

All models were trained using a supervised learning framework, with each patient sample classified as cancerous or normal. Classical machine learning models (Logistic Regression and SVM) were trained directly on fused multi-omics feature vectors. Backpropagation was used to train the MLP and Autoencoder models in order to minimize their loss functions. The autoencoder was trained unsupervised to rebuild input characteristics, and the encoded representations were then used for classification.

The proposed Transformer-based multi-omics fusion model was trained using supervised learning on the fused cfDNA-cfRNA dataset. The training set included 70% of the samples, with the remaining 30% retained for testing. To reduce overfitting, we used the AdamW optimizer with a learning rate of 0.001 and a weight decay of 1×10^{-4}.

The network was trained with a binary cross-entropy loss function and a batch size of 32. To improve convergence, the model was trained for 40-50 epochs with the learning rate gradually reduced using a step-based learning rate scheduler.

The Transformer architecture featured three Transformer encoder layers, each with four attention heads and a hidden dimension of 128 in the feed-forward network. The cfDNA and cfRNA feature vectors were first projected into 64-dimensional embeddings and then processed by the Transformer. The embedding and attention layers used dropout at 0.2-0.3 rates to improve their generalization performance. The selected hyperparameters enable the system to maintain optimal performance between high model complexity and effective cross-omics learning needed for cancer diagnosis.

3.5 Evaluation Metrics

The proposed model and baseline classifiers were evaluated through multiple standard evaluation criteria which provided thorough assessment of their performance. The measures evaluate two aspects of classification performance which include total accuracy and the ability to correctly identify cancer patients who require clinical treatment.

Accuracy measures the overall proportion of correctly identified samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision determines how many of the samples predicted as cancer were actually cancer cases.

$$Precision = \frac{TP}{TP + FP}$$

Recall (sensitivity) assesses the model's ability to correctly detect cancer patients:

$$Recall = \frac{TP}{TP + FN}$$

The F1-score is the harmonic mean of precision and recall, providing a balanced measure of categorization performance.

$$F1 - Score = 2 \times \frac{Precison \times Recall}{Precision + Recall}$$

In addition, the AUC-ROC curve was utilized to assess the models' overall discriminative ability across all decision thresholds. A greater AUC suggests better distinction of cancerous and normal samples.

The Precision-Recall curves were used to verify the behavior of each model when confronted with imbalance issues, providing a good understanding of the extent to which each model was able to maintain high accuracy while at the same time ensuring high sensitivity of cancer detection. The criteria discussed herein form a comprehensive evaluation of each model's cancer detection performance using multi-omics.

4. Result and Discussion

4.1 Comparative Classification Performance Across Models

The experimental results demonstrate that standard machine learning models and neural networks and the proposed Transformer-based multi-omics fusion technique exhibit major performance differences. Table 1 summarizes classification performance for all evaluation metrics.

Table 1. Classification performance of all models on the TCGA multi-omics

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.606	0.591	0.641	0.615	0.637
SVM	0.757	0.743	0.773	0.758	0.834
MLP	0.774	0.761	0.787	0.774	0.834
Autoencoder	0.539	0.526	0.614	0.566	0.576
Transformer (Ours)	**0.826**	**0.832**	**0.808**	**0.820**	**0.899**

The Transformer-based method shows better results than all other methods throughout every evaluation. The system achieves its highest performance with accuracy at 82.6% and precision at 83.2% and recall at 80.8% and F1-score at 82.0% and AUC at 0.899, which exceeds all baseline methods. The model successfully learns different molecular properties which enables it to perform pan-cancer detection at high efficiency. The model reaches better performance through self-attention because it enables the model to concentrate on important elements which exist in cfDNA and cfRNA samples.

The Transformer system delivers better performance through its ability to handle contextual information, which enables it to process data in multiple contextual situations. The gene expression signal determines how much weight the copy number effect should receive, while the reverse process also occurs. The process of dynamic balancing will produce increased sensitivity together with higher specificity, which the balanced precision and recall assessment confirms as evidence.

The F1-Score shows an exact value of 0.820. The system maintains strong performance because it efficiently manages both false positive and false negative errors. The system functions as an effective decision-making instrument which medical professionals can use at hospitals. The system demonstrates an AUC value of 0.899. The system demonstrates that its probability prediction method provides strong prediction results.

The accuracy of logistic regression analysis shows a low performance level which results in 60.6% accuracy and 0.637 AUC value. The data demonstrates that linear classifiers fail to identify the complex high-dimensional relationships which exist in the data.

The SVM model shows significant improvement, with 75.7% accuracy and 0.834 AUC. This shows that nonlinear kernel approaches can predict some interactions between genomic and transcriptome characteristics. However, SVMs rely on set kernel functions and are unable to dynamically adjust feature priority across modalities.

The MLP enhances performance (77.4% accuracy), demonstrating that neural networks can learn nonlinear representations from fused omics data. Nonetheless, its fully connected design treats all traits consistently and lacks means for explicitly capturing cross-omics relationships.

The autoencoder-based technique performs poorly, with only 53.9% accuracy. This finding demonstrates that unsupervised dimensionality reduction alone does not

ensure discriminative representations for cancer classification, particularly when the learnt embeddings are not specifically optimized for class separation.

4.2 Confusion Matrix Analysis of Model Predictions

Figure 2 illustrates the confusion matrix that shows the classification performance for all models This figure highlights the progressive improvement in classification quality from traditional machine learning to deep learning and Transformer-based fusion. The Transformer shows the lowest misclassification rates, confirming its ability to integrate cfDNA and cfRNA more effectively than baseline approaches.

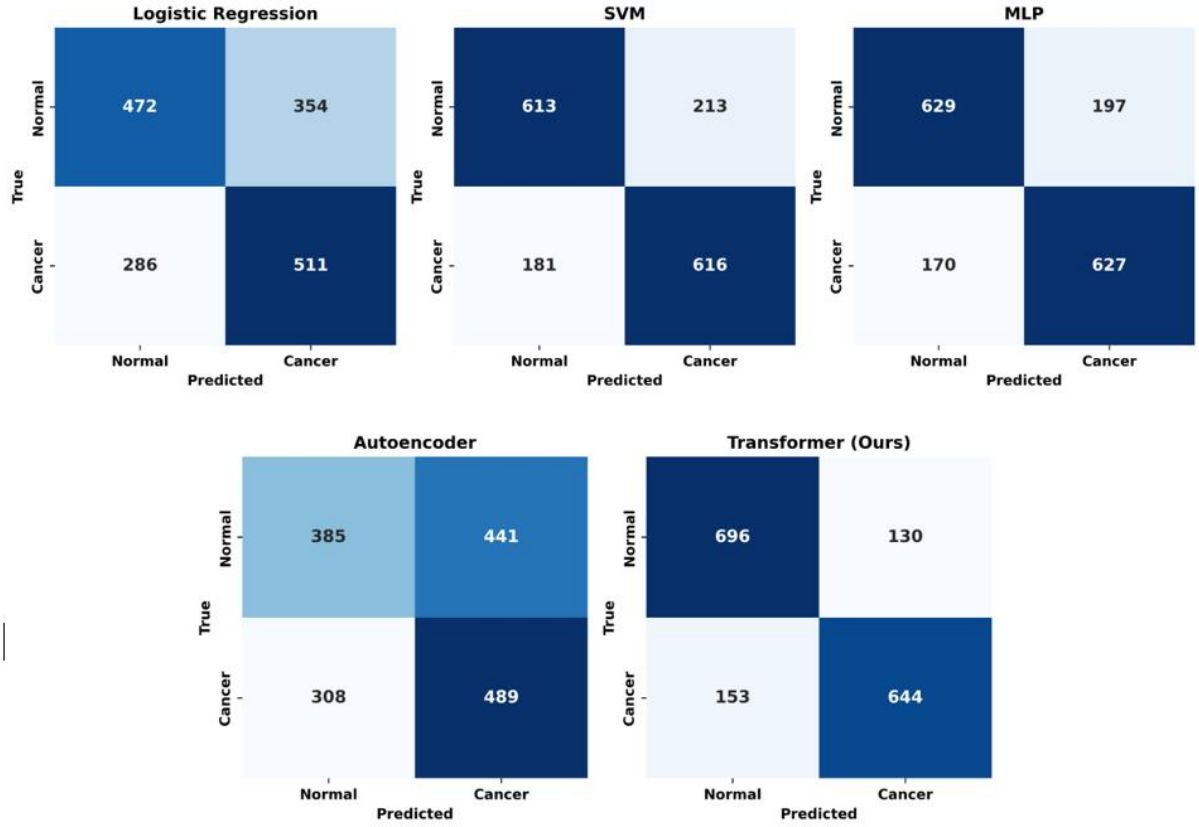

Figure 2. Confusion matrices of Logistic Regression, SVM, MLP, Autoencoder, and the proposed Transformer model.

4.3 ROC Curve Analysis and Discriminative Capability

The ROC curves show how true positive rates and false positive rates change at different decision thresholds. The recommended Transformer model achieves an area under the curve of 0.899 which enables it to differentiate between cancer samples and normal samples more effectively than other models. The ongoing curve growth demonstrates that the method produces better probability assessment results compared to traditional techniques. The ROC curve for all models which include Logistic Regression SVM MLP Autoencoder and the proposed Transformer-based model is displayed in Figure 3. So, our results present the ROC curve comparison of all models for multi-omics cancer detection. The Transformer model achieves the highest AUC (0.899), demonstrating superior discriminative ability. SVM and MLP show strong performance (AUC = 0.834), while Logistic Regression and Autoencoder perform comparatively lower, indicating limited capability in modeling complex cross-omics relationships.

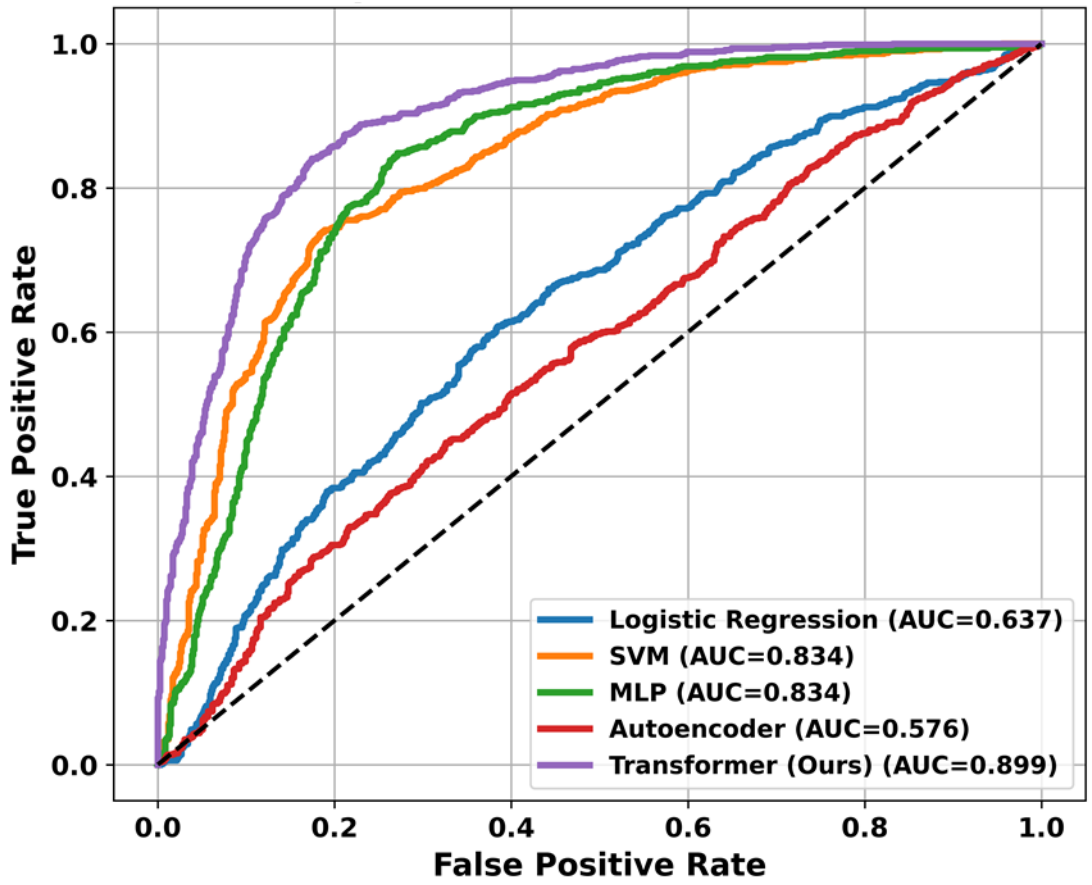

Figure 3. Receiver Operating Characteristic (ROC) curves for Logistic Regression, SVM, MLP, Autoencoder, and the proposed Transformer-based model.

4.4 Precision–Recall Curve Evaluation Under Class Imbalance

Precision-Recall curves provide information on model performance under class imbalance. The Transformer retains good precision and recall across thresholds, demonstrating that it accurately detects cancer patients while reducing false alarms. This makes the model ideal for clinical screening settings in which false positives and negatives are costly. Figure 4 illustrates the Precision-Recall curves for all models. Figure 4 illustrates the Precision–Recall curve comparison for all models on the TCGA multi-omics dataset. The Transformer model achieves the highest average precision (AP = 0.885), maintaining superior precision across nearly all recall levels. SVM (AP = 0.796) and MLP (AP = 0.773) demonstrate competitive but lower performance, while Logistic Regression (AP = 0.590) and Autoencoder (AP = 0.550) show limited effectiveness. The Transformer's ability to sustain high precision at increasing recall levels highlights its robustness under class imbalance, making it particularly suitable for clinical cancer screening scenarios where minimizing false negatives and false positives is critical.

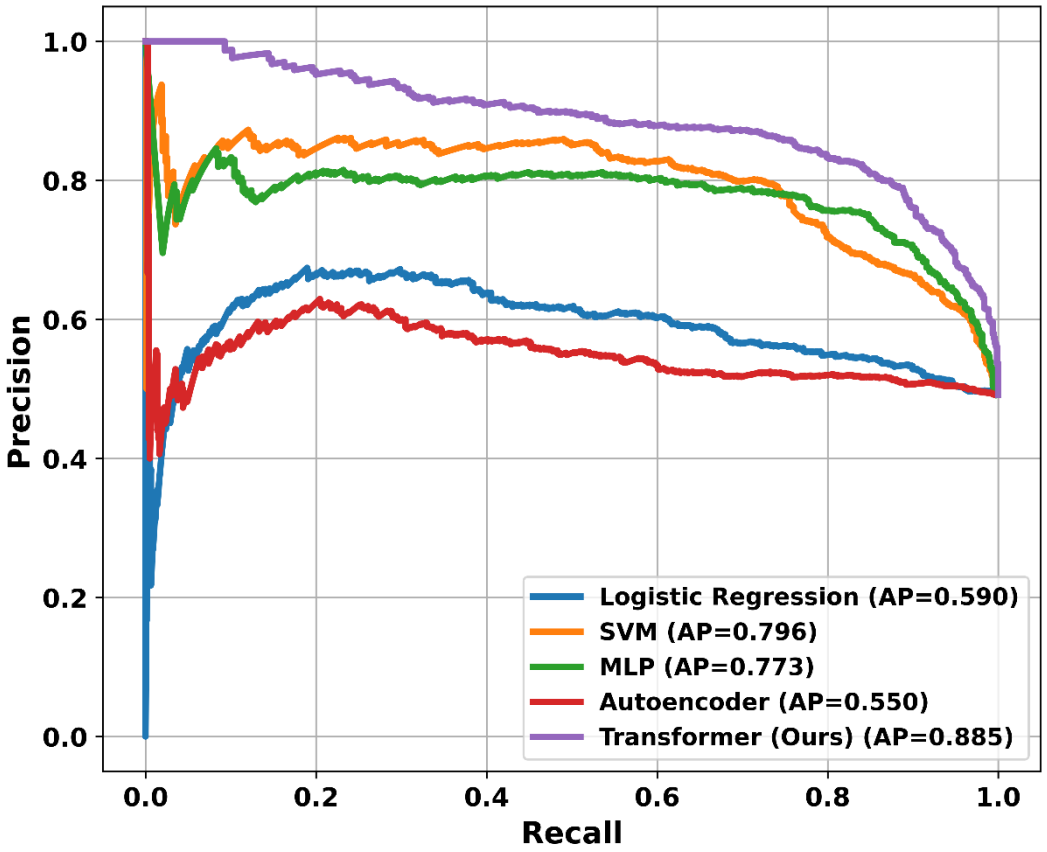

Figure 4. Precision–Recall curves comparing the cancer detection performance of all models on the TCGA dataset

4.5 Accuracy Improvement Across Increasing Model Complexity

Figure 5 depicts the improvement in total classification accuracy attained by increasingly advanced learning models. Traditional machine learning approaches, such as Logistic Regression and SVM, perform poorly because they are unable to describe complex nonlinear interactions in multi-omics data. Neural network-based models improve accuracy, but the suggested Transformer achieves the maximum accuracy, proving the efficacy of attention-based multi-omics fusion in cancer diagnosis. Figure 5 presents a bar chart comparing the classification accuracy of all evaluated models. The Transformer achieves the highest accuracy (82.6%), outperforming MLP (77.4%), SVM (75.7%), Logistic Regression (60.6%), and Autoencoder (53.9%). The progressive improvement from traditional machine learning to deep learning and attention-based architecture highlights the effectiveness of Transformer-driven multi-omics fusion. The results demonstrate that modeling cross-omics interactions significantly enhances cancer detection performance, validating the advantage of attention mechanisms over conventional feature concatenation approaches in complex, high-dimensional biological datasets.

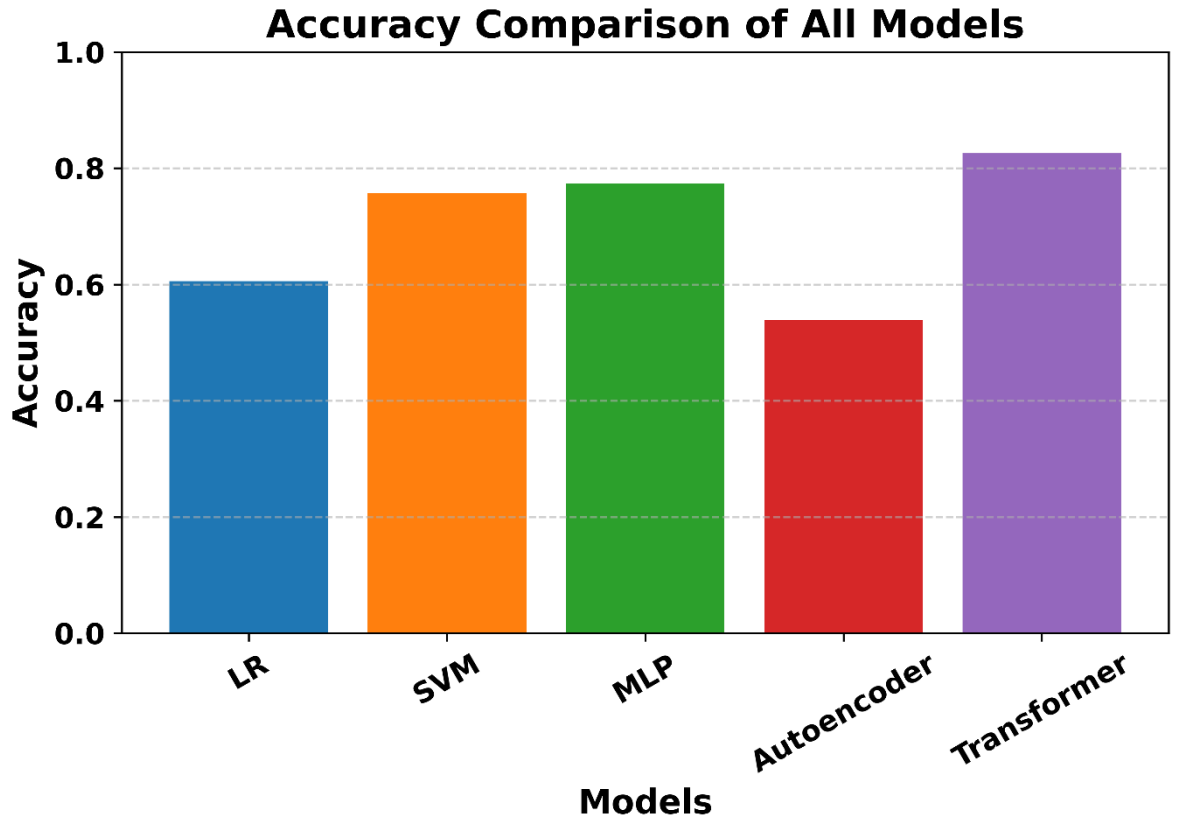

Figure 5. Bar chart comparing the classification accuracy of Logistic Regression (LR), SVM, MLP, Autoencoder, and the proposed Transformer-based multi-omics model on the TCGA dataset

4.6 Comprehensive Metric Comparison

Figure 6 provides a full comparison of categorization quality beyond accuracy. The suggested Transformer model consistently outperforms all baseline techniques in every metric, with the highest precision, recall, F1-score, and AUC. Strong memory indicates increased sensitivity for detecting cancer cases, and good precision predicts a reduced false-positive rate. The high AUC demonstrates that the Transformer generates well-calibrated probability estimates and superior class separation, making it more dependable for clinical decision support. Figure 6 presents a quantitative comparison of precision, recall, F1-score, and AUC across all models. The Transformer achieves the highest values in every metric (Precision = 0.832, Recall = 0.808, F1-score = 0.820, AUC = 0.899). SVM and MLP show moderate performance (AUC = 0.834), while Logistic Regression and Autoencoder perform substantially lower (AUC = 0.637 and 0.576, respectively). The results demonstrate that attention-based multi-omics fusion significantly improves classification effectiveness and overall discriminative power compared to conventional machine learning approaches.

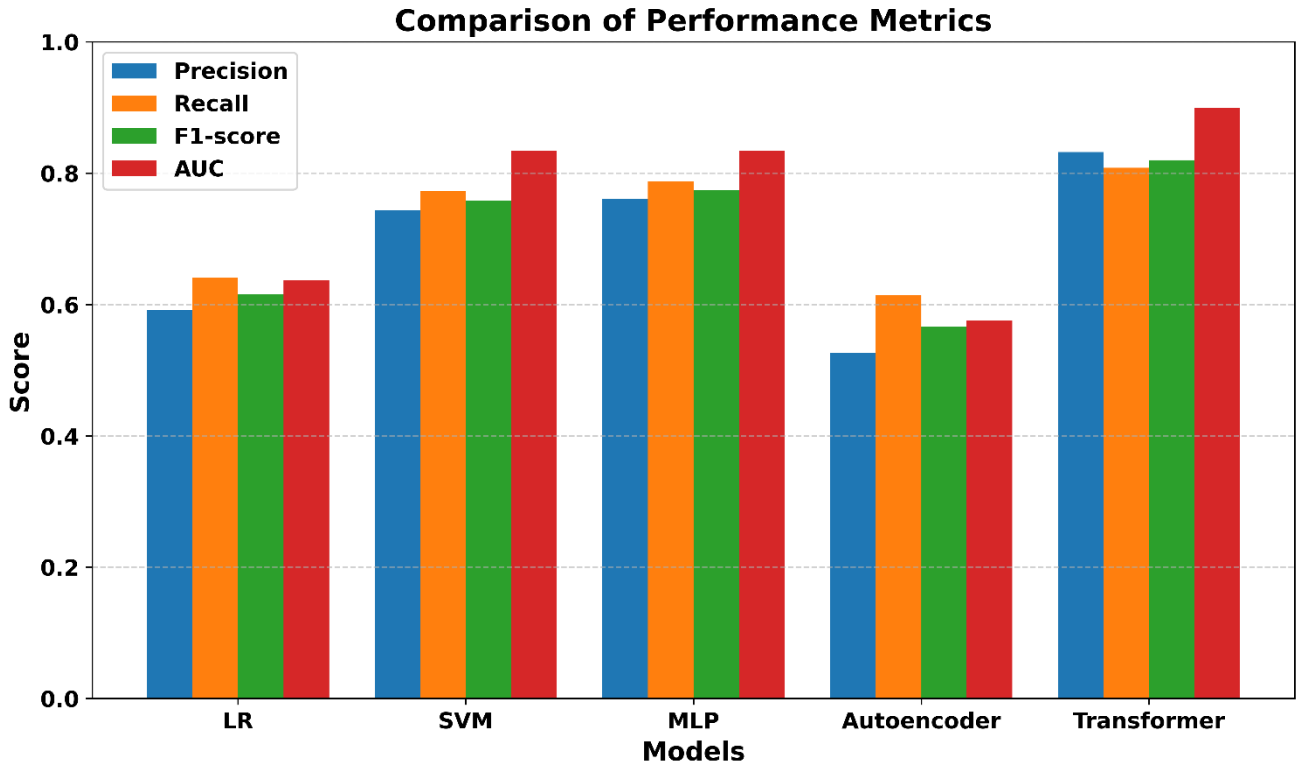

Figure 6. Performance comparison of all models across precision, recall, F1-score, and AUC

From a clinical standpoint, excellent recall is critical to reducing missed cancer diagnoses, and high precision decreases unneeded follow-up treatments. The suggested model meets both criteria, making it ideal for liquid biopsy-based early cancer diagnosis. Furthermore, the Transformer architecture is intrinsically scalable and may be expanded to include additional omics layers such as methylation and proteomics, increasing its usefulness in precision oncology.

5. Limitations and Future Work in U.S. Healthcare Systems

The proposed framework provides excellent performance yet presents multiple limitations. The model evaluation proceeded with testing for binary cancer detection yet the test did not include tumor subtype and tissue of origin identification. The study used TCGA dataset which contains extensive and diverse data yet researchers failed to use any external clinical validation datasets which results in limited ability to apply the findings to actual patient populations. The use of Transformer models demands higher computational resources compared to existing methods which creates challenges for their implementation in medical environments with limited resources (Ashik et al., 2023; Rahman et al., 2022).

The research will focus on three main areas to develop this existing framework through upcoming studies. The model will be extended to predict multi-class cancer subtypes and tissues of origin, allowing for more detailed diagnostic insights. The implementation of explainable AI mechanisms within the attention layers enables clinicians to trace which genomic regions and genes have the greatest impact on prediction outcomes. The research team will conduct independent liquid biopsy cohort validation together with actual clinical data assessment to establish system robustness and dependable functioning and their ability to be implemented in real-world situations.

5.1 Limitations

The only data used in this study are the multi-omics datasets from the Translational Cancer Genome Atlas (TCGA). These datasets, while extensive and diverse, are retrospective and primarily derived from tissue-based tumor samples rather than authentic liquid biopsy cohorts. Recent assessments of liquid biopsy indicate that clinical implementation requires validation via independent plasma datasets collected prospectively to ensure generalizability across a diverse population (Zhu et al., 2026; Abreu et al., 2026). It is still unclear how well the model works in

different clinical settings without outside validation. Moreover, studies aimed at the early detection of various cancers, exemplified by the PROMISE trial, underscore the necessity of executing thorough prospective validation cohorts before clinical application (Duan et al., 2026). Even though TCGA has a good method, it's still important to try it out in the real world. The main goal of the current paradigm is to tell the difference between cancerous and healthy tissue. But you also need to know the subtype of the tumor and the tissue it came from (TOO) to make clinical screening apps. Luo et al. (2025) and Kwon et al. (2023) assert that multi-omics methodologies exhibit superior efficacy in tissue-specific classification by amalgamating DNA methylation, mutation patterns, and transcriptome data. If the Transformer design could also predict different types of cancer, it would work much better as a treatment.

But the model doesn't have any extra molecular layers, like DNA methylation, fragmentomics, proteomics, or epigenomic markers. This is true even though it changes how many copies of cfDNA there are and how cfRNA genes are expressed. Liang et al. (2025) and Song et al. (2025) assert that the integration of multiple methodologies facilitates the early detection of cancer. Research on cfDNA multi-omics profiling has shown that combining methylation and fragmentation signals greatly improves the accuracy of early detection and classification (Song et al., 2025).

Adding more omics layers to future architectures is a good idea because it will make the system easier to understand and better at predicting what will happen in biology. Transformer designs work well, but they need a lot more processing power than regular machine learning algorithms. Recent assessments of deep learning for multi-omics integration reveal that attention-based models require enhanced memory capacity, graphics processing unit (GPU) resources, and extended training periods (Sartori et al., 2025; Baião et al., 2025). This might make it hard to use in hospitals that don't have a lot of money or labs that don't have a lot of people. Model pruning, knowledge distillation, and lightweight transformer versions are some tools that might help get around these limits.

The current version doesn't have any formal explainable artificial intelligence (XAI) modules, but the attention method does help make some things clearer. In precision oncology, clinical acceptance hinges on the capacity to predict outcomes and the proficiency in analyzing biological processes. Deep learning multi-omics models must integrate interpretability methodologies to discern significant genes, CNV regions, or transcriptome signatures that impact predictions (Pushparaj & Muthukumar, 2026; Sartori et al., 2025).

5.2 Future Directions

Furthermore, it is widely known that the utilization of plasma-derived cfDNA and cfRNA datasets for the purpose of independent validation is of utmost significance. It has been suggested that multi-center prospective studies that are comparable to the PROMISE research architecture (Duan et al., 2026) be utilized in order to test the robustness, repeatability, as well as the sensitivity and specificity at the population level. This is in order to determine whether or not the research in question is accurate. This is due to the fact that these investigations demonstrate similarities to the research framework known as PROMISE. A Multi-Omics Transformer Foundation Model has the potential to transform early cancer detection in the United States by leveraging large-scale data analytics and AI-driven decision-making frameworks, as emphasized in recent studies on data-driven healthcare and governance (Ashik et al., 2023; Hossain et al., 2024). Drawing on insights from big data applications in economic planning, policy analytics, and digital transformation, such a model could enhance precision oncology, optimize resource allocation, and improve cost-efficiency across U.S. healthcare systems (Islam et al., 2023; Khan et al., 2024). In the future, this approach may support proactive early intervention strategies, strengthen healthcare sustainability, and align with national priorities for technology-enabled, value-based care.

Future research should expand the integration of transformer-based multi-omics architectures with real-world clinical datasets to enhance predictive robustness and generalizability across diverse U.S. populations. Building upon recent work demonstrating the potential of machine learning in cancer prediction and disease pattern recognition (Rishad et al., 2025), future models should incorporate longitudinal cfDNA and cfRNA sequencing data, epigenomic signatures, and federated learning frameworks to ensure privacy-preserving, cross-institutional validation. Moreover, scalable AI frameworks aligned with healthcare system optimization strategies (Sufian et al., 2024) could facilitate deployment within U.S. oncology workflows, enabling early-stage tumor detection, risk stratification, and personalized therapeutic guidance. Integrating explainable AI mechanisms will also be essential to ensure regulatory compliance, clinician trust, and equitable access in precision oncology ecosystems.

When the design is expanded to include the identification of several cancer subtypes, there is a possibility that the clinical screening applicability potential may increase. This is a possibility. The use of transformer-based sequence modeling has been shown to be beneficial in a variety of various areas of biological prediction, as stated by Zhou et al. (2025). This was previously demonstrated. For the purpose of achieving the objective of obtaining diagnostic insights that can be put into practice, the introduction of tissue-of-origin categorization into the attention process can be useful.

When proteomics, epigenomic fragmentation profiles, and methylation signatures are incorporated into the detection process, there is a possibility that the accuracy of the detection might be significantly increased. According to the findings of research that was conducted by Ran et al. (2025), multi-omics artificial intelligence systems that combine genomic and transcriptomic layers have demonstrated advancements in the classification of subtypes and the prediction of treatment responses. Increasing the number of omics modalities beyond two would result in an improvement in both the clinical dependability of the study as well as the biological depth of the investigation. Within the context of gaining a knowledge of the biologically significant drivers of predictions, it is of the utmost importance to combine attention-weight visualization, SHAP analysis, and gene-level attribution mapping. According to Pushparaj and Muthukumar (2026), this would be a significant step toward bridging the gap between computer modeling and the decision-making process in precision oncology. They have stated that this would be a significant success. As emphasized by Guria et al. (2025), artificial intelligence has the potential to bridge structural disparities and promote inclusive development across healthcare systems. Therefore, future transformer foundation models should incorporate fairness-aware training mechanisms and population-stratified validation pipelines to minimize algorithmic bias in early cancer detection.

It is suggested that the work that will be done in the future should concentrate on the creation of lightweight transformer designs, federated learning models for cross-institution collaboration, and secure cloud-based deployment frameworks (Rahman et al., 2025). These are the areas that should be prioritized. It is anticipated that the deployment of these modifications will lead to enhancements in scalability as well as accessibility across all healthcare systems (Tanvir et al., 2020).

6. Conclusion

The paper introduced a multi-omics fusion framework which uses Transformer technology to diagnose all cancer types through cfDNA and cfRNA analysis of TCGA data. The model used an attention-based architecture which processed gene expression data together with copy number variation information to model complex biological relationships that standard machine learning and basic neural networks could not grasp. The experimental results demonstrated that the proposed model achieved better performance results than Logistic Regression SVM MLP and autoencoder-based approaches across all evaluation metrics. The Transformer demonstrated its ability to distinguish between cancer and normal samples through its highest accuracy of 82.6% and F1-score of 82.0% and AUC value of 0.899. The results demonstrate how attention-based multi-omics learning helps achieve precise and efficient cancer detection through liquid biopsy methods.

Funding

This research received no external funding.

Conflicts of Interest

No potential conflict of interest was reported.

References

1. Abreu, R. da S., Ferreira, D. D. P., Araujo, N. S. de, Horita, S., Tilli, T. M., Degrave, W., Moreira, A. dos S., & Waghabi, M. C. (2026). Liquid biopsy in cancer diagnosis and prognosis: A paradigm shift in precision oncology. Frontiers in Molecular Biosciences, 12, Article 1708518. https://doi.org/10.3389/fmolb.2025.1708518
2. Ana R Baião, Zhaoxiang Cai, Rebecca C Poulos, Phillip J Robinson, Roger R Reddel, Qing Zhong, Susana Vinga, Emanuel Gonçalves, A technical review of multi-omics data integration methods: from classical statistical to deep generative approaches, Briefings in Bioinformatics, Volume 26, Issue 4, July 2025, bbaf355, https://doi.org/10.1093/bib/bbaf355
3. Ashik, A. A. M., Rahman, M. M., Hossain, E., Rahman, M. S., Islam, S., & Khan, S. I. (2023). Transforming U.S. Healthcare Profitability through Data-Driven Decision Making: Applications, Challenges, and Future Directions. European Journal of Medical and Health Research, 1(3), 116-125. https://doi.org/10.59324/ejmhr.2023.1(3).21
4. Baião, A.R., Cai, Z., Poulos, R.C., Robinson, P.J., Reddel, R.R., Zhong, Q., Vinga, S. and Gonçalves, E.,

2025. A technical review of multi-omics data integration methods: from classical statistical to deep generative approaches. Briefings in bioinformatics, 26(4), p.bbaf355.

5. Duan, J., Gao, Q., Wang, Z., Cai, S., Fan, J., Wang, J., et al. (2026). Exploration of multi-omics liquid biopsy approaches for multi-cancer early detection: The PROMISE study. Journal of Molecular Diagnostics, 7(1), 100176. https://doi.org/10.1016/j.xinn.2025.101076
6. Gonzalez Reymundez, A. (2020) Multi-omic pan-cancer data from TCGA. Mendeley Data, V2. doi:10.17632/r8p67nfjc8.2.
7. Guria, Z. M., Morshed, N., Rahman, I., Dhar, S. R., & Sufian, M. A. (2025). Advancing global peace through inclusive economic development and the role of artificial intelligence in bridging socioeconomic divides. In Proceedings of the 2025 International Conference on Artificial Intelligence's Future Implementations (ICAIFI) (pp. 100–105). IEEE. https://doi.org/10.1109/ICAIFI66942.2025.11326547
8. Hossain, E., Ashik, A. A. M., Rahman, M. M., Khan, S. I., Rahman, M. S., & Islam, S. (2023). Big data and migration forecasting: Predictive insights into displacement patterns triggered by climate change and armed conflict. Journal of Computer Science and Technology Studies, 5(4): 265–274. https://doi.org/10.32996/jcsts.2023.5.4.27/.
9. Hossain, E., Shital, K. P., Rahman, M. S., Islam, S., Khan, S. I., & Ashik, A. A. M. (2024). Machine learning-driven governance: Predicting the effectiveness of international trade policies through policy and governance analytics. Journal of Trends in Financial and Economics, 1(3), 50–62. https://doi.org/10.61784/jtfe3053. https://scholar.google.com/citations?view_op=view_citation&hl=en&user=iOJQX0sAAAAJ&sortby=pubdate&citation_for_view=iOJQX0sAAAAJ:4TOpqqG69KYC
10. Islam, S., Hossain, E., Rahman, M. S., Rahman, M. M., Khan, S. I., & Ashik, A. A. M. (2023). Digital Transformation in SMEs: Unlocking Competitive Advantage through Business Intelligence and Data Analytics Adoption. 5 (6):177-186. https://doi.org/10.32996/jbms.2023.5.6.14
11. Islam, S., Khan, S. I., Ashik, A. A. M., Hossain, E., Rahman, M. M., & Rahman, M. S. (2024). Big data in economic recovery: A policy-oriented study on data analytics for crisis management and growth planning. Journal of Computational Analysis and Applications (JoCAAA), 33(7), 2349–2367. Retrieved from https://www.eudoxuspress.com/index.php/pub/article/view/3338
12. Juie, B. J. A., Kabir, J. U. Z., Ahmed, R. A., & Rahman, M. M. (2021). Evaluating the impact of telemedicine through analytics: Lessons learned from the COVID-19 era. Journal of Medical and Health Studies, 2(2), 161–174.https://doi.org/10.32996/jmhs.2021.2.2.19
13. Khan, S. I., Rahman, M. S., Ashik, A. A. M., Islam, S., Rahman, M. M., & Hossain, E. (2024). Big Data and Business Intelligence for Supply Chain Sustainability: Risk Mitigation and Green Optimization in the Digital Era. European Journal of Management, Economics and Business, 1(3): 262-276. https://doi.org/10.59324/ejmeb.2024.1(3).23
14. Kwon, H.-J., Park, U.-H., Goh, C. J., Park, D., Lim, Y. G., Lee, I. K., Do, W.-J., Lee, K. J., Kim, H., Yun, S.-Y., Joo, J., Min, N. Y., Lee, S., Um, S.-W., & Lee, M.-S. (2023). Enhancing Lung Cancer Classification through Integration of Liquid Biopsy Multi-Omics Data with Machine Learning Techniques. Cancers, 15(18), 4556. https://doi.org/10.3390/cancers15184556
15. Liang X, Tang Q, Chen J, Wei Y. Liquid Biopsy: A Breakthrough Technology in Early Cancer Screening. Cancer Screen Prev. 2025;4(1):40-52. doi: 10.14218/CSP.2024.00031.
16. Luo X, Xie S, Hong F, Li X, Wei Y, Zhou Y, Su W, Yang Y, Tang L, Dao F, Cai P, Lin H, Lai H, Lyu H. From multi-omics to deep learning: advances in cfDNA-based liquid biopsy for multi-cancer screening. Biomark Res. 2025 Nov 28;14(1):3. doi: 10.1186/s40364-025-00874-z.
17. Mohib, M. M., Uddin, M. B., Rahman, M. M., Tirumalasetty, M. B., Al-Amin, M. M., Shimu, S. J., Alam, M. F., Arbee, S., Munmun, A. R., Akhtar, A., & Mohiuddin, M. S. (2025). Dysregulated Oxidative Stress Pathways in Schizophrenia: Integrating Single-Cell Transcriptomic and Human Biomarker Evidence. Psychiatry International, 6(3), 104. ttps://doi.org/10.3390/psychiatryint6030104
18. Pushparaj, A. K., & Muthukumar, M. (2026). Deep Learning Architectures for Multi-Omics Data Integration: Bridging Biomarker Discovery and Clinical Translation. Preprints. https://doi.org/10.20944/preprints202601.1884.v1
19. Rahman, M. M., Juie, B. J. A., Tisha, N. T., & Tanvir, A. (2022). Harnessing predictive analytics and machine learning in drug discovery, disease

surveillance, and fungal research. Eurasia Journal of Science and Technology, 4(2), 28-35. https://doi.org/10.61784/ejst3099

20. Rahman, M. M., Rahman, M. S., Islam, S., Khan, S. I., Ashik, A. A. M., Hossain, E., & Tanvir, A. (2025). Integrating data analytics into health informatics: Advancing equity, pharmaceutical outcomes, and public health decision-making. Eurasian Journal of Medicine and Oncology, 9(4), 284–295. https://doi.org/10.36922/EJMO025300319
21. Ran, D., Li, J., Zhao, M., Du, L., Zhang, Y., & Zhu, J. (2025). Artificial intelligence integrates multi-omics data for precision stratification and drug resistance prediction in breast cancer. Frontiers in Oncology, 15, Article 1612474. https://doi.org/10.3389/fonc.2025.1612474.
22. Rishad, S. S. I., Akter, N., Ahamed, A., Sufian, M. A., Rinky, A. I., & Rimi, N. N. (2025). Leveraging Machine Learning for Cancer Insights and Disease Predictions. In Proceedings of the 2025 IEEE International Conference on Data-Driven Social Change (ICDDSC-2025), Pakistan. IEEE.
23. Sartori, F., Codicè, F., Caranzano, I., Rollo, C., Birolo, G., Fariselli, P., & Pancotti, C. (2025). A Comprehensive Review of Deep Learning Applications with Multi-Omics Data in Cancer Research. Genes, 16(6), 648. https://doi.org/10.3390/genes16060648
24. Song S, Zhang X, Cui P, He W, Zhou J, Wang S, Xiong Y, Xu S, Lin X, Huang G, Tan X, Xu Q, Liu Y, Li Q, Yuan K, Feng M, Lai H, Yang H, Zhang S. Plasma cfDNA multi-omic biomarkers profiling for detection and stratification of gastric carcinoma. BMC Cancer. 2025;25(1):1003. doi: 10.1186/s12885-025-14409-0.
25. Sufian, M. A., Rimon, S. M. T. H., Mosaddeque, A. I., Guria, Z. M., Morshed, N., & Ahamed, A. (2024). Leveraging machine learning for strategic business gains in the healthcare sector. In Proceedings of the 2024 International Conference on TVET Excellence & Development (ICTeD) (pp. 225–230). IEEE. https://doi.org/10.1109/ICTeD62334.2024.10844658
26. Tanvir, A., Juie, B. J. A., Tisha, N. T., & Rahman, M. M. (2020). Synergizing big data and biotechnology for innovation in healthcare, pharmaceutical development, and fungal research. International Journal of Biological, Physical and Chemical Studies, 2(2), 23–32. https://doi.org/10.32996/ijbpcs.2020.2.2.4.
27. Tanvir, A.; Jo, J.; Park, S.M. (2024). Targeting Glucose Metabolism: A Novel Therapeutic Approach for Parkinson's Disease. Cells. 13, 1876. https://doi.org/10.3390/cells13221876
28. Zhong, P., Bai, L., Hong, M., Ouyang, J., Wang, R., Zhang, X., & Chen, P. (2024). A Comprehensive Review on Circulating cfRNA in Plasma: Implications for Disease Diagnosis and Beyond. Diagnostics, 14(10), 1045. https://doi.org/10.3390/diagnostics14101045
29. Zhou, S., Guan, C., Deng, S. et al. A novel sequence-based transformer model architecture for integrating multi-omics data in preterm birth risk prediction. npj Digit. Med. 8, 536 (2025). https://doi.org/10.1038/s41746-025-01942-2
30. Zhu, H., Li, Z., Xie, K., Kassim, S. H., Cao, C., Huang, K., Lu, Z., Ma, C., Li, Y., Jiang, K., & Yin, L. (2026). Liquid Biopsy in Early Screening of Cancers: Emerging Technologies and New Prospects. Biomedicines, 14(1), 158. https://doi.org/10.3390/biomedicines14010158