A Machine Learning Framework for Tumour Classification Using Transcriptomic and Multi-Omics Datasets

Rhea Maria James
Computer Science Engineering
Amal Jyothi College of Engineering
Kanjirappally, India
rheamariajames2026@cs.ajce.in

Richy Sara George
Computer Science Engineering
Amal Jyothi College of Engineering
Kanjirappally, India
richysarageorge2026@cs.ajce.in

Sayooj Kumar M
Computer Science Engineering
Amal Jyothi College of Engineering
Kanjirappally, India
sayoojkumarm2026@cs.ajce.in

Nihal Muhammed Ayoob
Computer Science Engineering
Amal Jyothi College of Engineering
Kanjirappally, India
nihalmuhammedayoob2026@cs.ajce.in

Shan Krishna
Computer Science Engineering
Amal Jyothi College of Engineering
Kanjirappally, India
shankrishna2026@cs.ajce.in

Asst. Prof. Tintu Alphonsa Thomas
Computer Science Engineering
Amal Jyothi College of Engineering
Kanjirappally, India
tintualphonsathomas@amaljyothi.ac.in

***Abstract*—Cancer is a biologically heterogeneous disease characterized by molecular alterations across multiple regulatory layers, necessitating robust computational modelling for accurate diagnosis and biomarker discovery. The increasing availability of high-dimensional genomic and multi-omics datasets from large-scale initiatives such as The Cancer Genome Atlas (TCGA) has enabled the development of machine learning approaches for cancer classification. However, challenges including extreme dimensionality, feature redundancy, and class imbalance continue to affect model stability and generalization performance.**

In this study, we propose a reproducible integrative machine learning framework for tumor versus normal classification and biomarker identification using gene expression and multi-omics TCGA data. The methodology employs Extreme Gradient Boosting (XGBoost) for embedded feature selection to identify the most informative molecular variables from tens of thousands of features. The selected features are subsequently used to train ensemble classifiers including Logistic Regression, Random Forest, and Support Vector Machine models.

To ensure unbiased performance estimation and prevent data leakage, a stratified five-fold cross-validation strategy is adopted. Experimental evaluation on breast and lung cancer datasets demonstrates strong discriminative performance, with the XGBoost–Random Forest model achieving mean classification accuracies exceeding 99%, along with high ROC-AUC and Cohen's Kappa values. Furthermore, multi-omics integration improves classification robustness by capturing complementary molecular signals across biological layers.

The results indicate that XGBoost-driven feature selection combined with ensemble learning provides a scalable, interpretable, and effective framework for high-dimensional cancer classification and biomarker discovery.

***Index Terms*—Cancer classification, Machine learning, Multi-omics integration, XGBoost, Biomarker discovery.**

I. INTRODUCTION

Cancer is a wide category of diseases that are defined by the uncontrolled growth of cells due to genetic and molecular changes. Malignant cells grow rapidly, infiltrate neighbouring tissues, and can also travel to other organs, causing the failure of the targeted organs and resulting in death. Cancer is the second leading cause of death after cardiovascular diseases and is one of the most threatening health issues in the world [1]. The threat of cancer is further increased by its biological diversity, late diagnosis, and resistance to treatment, making early and accurate detection of utmost importance.

The incidence of cancer is also a serious issue in developing countries like India. According to epidemiological studies, one in every nine people in India is likely to suffer from cancer in their lifetime, and breast cancer is the most common type of cancer in women, followed by lung cancer in men [2]. According to the Global Cancer Observatory, India is the third country in the world in terms of cancer incidence in 2020, and estimates show a 57.5% rise in cancer incidence by 2040.

The increasing volume of biomedical data has led to the emergence of artificial intelligence (AI) and machine learning (ML) as revolutionary technologies in the field of oncology. Machine learning, a branch of AI, allows computers to learn from patterns in large biomedical datasets and make predictions without being explicitly programmed [3]. Unlike traditional statistical modeling, ML algorithms are capable of processing noisy, high-dimensional, and nonlinear biomedical data. Various studies have shown that AI-based systems can provide diagnostic accuracy comparable to, if not superior to, that of expert human clinicians in the fields of breast cancer screening, prostate cancer diagnosis, and dermatological cancer diagnosis [4]–[8]. Such achievements have led to an explosion of research in the use of AI for cancer risk assessment, diagnosis, prognosis, and treatment decision-making.

However, many existing cancer classification models have

been shown to depend largely on single-omics data, especially gene expression profiles. While single-omics models have demonstrated high classification accuracy, they have offered a restricted and partial perspective on the molecular processes that underlie cancer [1]. Cancer is, in fact, a multi-faceted disease that is driven by genomic mutations, epigenetics, transcriptional aberrations, proteomic changes, and metabolic dysregulation. Single-omics models are, therefore, prone to noise, batch effects, and tumour heterogeneity.

To address the above limitations, multi-omics has recently been recognized as a highly successful paradigm in cancer research. Multi-omics is a strategy that combines data from multiple biological platforms, such as genomics, transcriptomics, epigenomics, proteomics, and metabolomics, to offer a holistic understanding of tumour biology [12], [13]. Large-scale projects such as The Cancer Genome Atlas (TCGA) have made it possible to perform multi-omics profiling of various types of cancers, thereby facilitating multi-omics analysis and improved disease understanding. Multi-omics models have the ability to capture interactions among multiple biological platforms, thereby offering improved predictive accuracy, better biological interpretability, and improved robustness compared to single-omics models.

However, the integration of multi-omics data also poses substantial computational challenges in terms of its extreme dimensionality, diversity, and complex interdependencies among different data types. Hence, the need for advanced machine learning algorithms that can efficiently perform feature selection, dimensionality reduction, and accurate classification is paramount. In this regard, ensemble learning algorithms such as Extreme Gradient Boosting (XGBoost), when paired with linear or margin-based classifiers, provide a promising approach to developing accurate and interpretable models for cancer prediction.

Inspired by these challenges, this research work proposes a multi-omics cancer classification model based on TCGA breast cancer datasets and advanced machine learning algorithms. The proposed model assesses the hybrid strategy of combining XGBoost with Logistic Regression and Support Vector Machines, with strict stratified cross-validation to evaluate model performance. The significance of this research work is embedded in its emphasis on developing robust generalization, biological interpretability, and clinical applicability, thus adding to the rapidly expanding area of AI-powered precision oncology.

The primary contributions of this study are summarized as follows:

- **Reproducible Machine Learning Framework:** We propose a reproducible machine learning framework for cancer classification using high-dimensional transcriptomic and multi-omics datasets derived from TCGA.
- **Embedded Feature Selection Strategy:** We implement an XGBoost-based embedded feature ranking approach to identify the most informative molecular features, substantially reducing dimensionality while preserving predictive performance.
- **Hybrid Classification Architecture:** We evaluate hybrid classification models that combine XGBoost-driven feature selection with Logistic Regression, Random Forest, and Support Vector Machine classifiers under stratified cross-validation to ensure unbiased and stable performance estimation.
- **Comparative Multi-Omics Evaluation:** We conduct a comparative analysis across transcriptomic-only and multi-omics datasets to assess model robustness, performance consistency, and the impact of integrative molecular representation.
- **Biomarker Identification and Interpretability:** We identify biologically relevant candidate genes associated with tumour progression, supporting the interpretability and potential translational relevance of the proposed framework in precision oncology.

II. Motivation and Objectives

A. Motivation

Cancer is still one of the most challenging and life-threatening diseases because of its molecular diversity, aggressive nature, and variability in treatment response. Despite major breakthroughs in biomedical research, early diagnosis and precise classification of cancer are still one of the biggest challenges in clinical practice [2]. These challenges create a pressing need for the development of accurate, automated, and scalable diagnostic tools that can help clinicians in early cancer diagnosis and decision-making.

Recent breakthroughs in artificial intelligence and machine learning have shown immense promise in cancer research, as they have made it possible to analyze complex high-dimensional biomedical data [3], [4]. Machine learning algorithms have shown expert-level performance in various cancer-related tasks such as breast cancer screening and prostate cancer diagnosis, which indicates their potential in identifying hidden patterns that are beyond human observation [7], [8]. However, most of the existing AI-based cancer classification models are based on single-omics datasets, which represent a very limited aspect of cancer biology and often lack generalizability across datasets because of noise, batch effects, and biological variability [1].

Cancer is a multi-factorial disease that depends upon complex interactions at multiple levels of molecular complexity, such as genomic mutations, epigenetics, transcriptional modifications, and downstream protein and metabolic changes. By themselves, single-omics studies are inadequate to capture this complexity [12]. Multi-omics integration offers a more holistic understanding of cancer biology by simultaneously analyzing multiple data sources, allowing for better biomarker identification, improved classification accuracy, and better robustness in predictive modelling [13]. Large-scale projects like The Cancer Genome Atlas (TCGA) have made high-quality multi-omics data publicly available, offering unprecedented opportunities for data-driven, integrative cancer research.

However, the application of multi-omics approaches is impeded by challenges associated with extreme dimensionality,

data heterogeneity, and the need for efficient feature selection and model generalization. The need for sophisticated machine learning infrastructure that can effectively integrate multi-omics data while being interpretable and meaningful from a clinical perspective is, therefore, critical. These challenges represent the central motivation for this research.

B. Objectives

The primary objective of this project is to develop a robust and generalizable cancer classification framework using multi-omics data and advanced machine learning techniques. Specifically, the objectives of this study are as follows:

- To develop a multi-omics cancer classification model using publicly available TCGA cancer datasets, integrating high-dimensional molecular features for better disease modelling.
- To explore the utility of ensemble machine learning algorithms, specifically Extreme Gradient Boosting (XGBoost), for feature selection and representation learning in multi-omics data.
- To develop and compare hybrid cancer classification models combining XGBoost with Logistic Regression, Support Vector Machines, and Random Forest algorithms, and to assess their predictive performance using strict cross-validation.
- To explore biologically meaningful molecular features underlying cancer prediction, thereby improving model interpretability and facilitating potential clinical inferences.

By achieving these objectives, this work aims to contribute to the advancement of AI-driven precision oncology by providing a scalable, interpretable, and data-driven framework for cancer classification using multi-omics information.

III. Literature Survey

The tremendous progress made in high-throughput sequencing technology has resulted in an unprecedented increase in genomic and molecular data, thereby triggering intense research on machine learning-based cancer classification. The existing research works can be generally classified into single-omics learning methods, pan-cancer classification frameworks, ensemble machine learning models, explainable AI-based biomarker identification, and multi-omics integration strategies.

A. Machine Learning for Cancer Classification Using Single-Omics Data

The initial research works on cancer classification were based on single-omics data, specifically gene expression data obtained from microarray and RNA-Seq analyses. Alharbi and Vakanski [1] provided a thorough review of machine learning methods used for gene expression-based cancer classification, pointing out the extensive use of Support Vector Machines (SVM), Random Forests (RF), k-Nearest Neighbors (kNN), and neural networks. Although these methods showed high classification accuracy in a controlled environment, they tended to perform poorly when used across multiple datasets due to biological heterogeneity and batch effects.

Some studies have shown robust results using conventional machine learning algorithms. Hoadley et al. [14] tested various classifiers on TCGA RNA-Seq data and showed an accuracy of 95.8% using linear SVMs with variance-driven feature selection. Likewise, Li et al. [22] designed a genetic algorithm-k-nearest neighbor (GA-kNN) classifier for cancer diagnosis, which showed an overall accuracy of 90%, although there was a compromise on the performance for less common types of cancers.

Later, deep learning algorithms, such as convolutional neural networks (CNNs), were developed to capture complex patterns in gene expression data [23], [24]. Although these algorithms showed high overall predictive accuracy, there was a significant misclassification rate for cancers arising from biologically similar tissues, indicating the inability of these models to capture the minute molecular differences among these tissues.

Collectively, these findings suggest that single-omics approaches, although effective in capturing transcriptomic variation, provide an incomplete representation of cancer biology and may be insufficient to model cross-layer molecular interactions driving tumour progression.

B. XGBoost-Based Models for Cancer Prediction

Ensemble learning methods, particularly Extreme Gradient Boosting (XGBoost), have demonstrated strong performance in genomic data analysis due to their ability to capture non-linear feature interactions and handle high-dimensional data efficiently. [10] systematically evaluated multiple machine learning and deep learning models for pan-cancer classification using genomic alterations and found XGBoost to consistently outperform other classifiers in terms of accuracy and robustness. Their study emphasized XGBoost's balance between expressive power and model simplicity.

More recently, Ghuriani et al. [9] proposed the XGB-BIF framework, which integrates XGBoost-based feature selection. Their approach achieved classification accuracies exceeding 90% and demonstrated strong agreement using Cohen's Kappa statistic. Notably, XGB-BIF discovered biologically relevant biomarkers for gastric, breast, and lung cancers and externally validated its performance on the METABRIC dataset, which emphasizes the translational value of XGBoost-based frameworks.

The above researches encourage the application of XGBoost not only as a predictive model but also as a robust feature selection tool for biomarker discovery in high-dimensional genomic data.

C. Multi-Omics Integration for Cancer Classification

To address the shortcomings of single-omics models, the concept of multi-omics integration has recently gained immense popularity in the field of cancer research. Argelaguet et al. [11] proposed the concept of Multi-Omics Factor Analysis (MOFA), an unsupervised learning method that identifies

common and modality-specific factors of variation in multi-omics data. MOFA has been successfully used for the analysis of cancer cohorts to discover hidden subtypes of disease and underlying molecular mechanisms.

Extensive reviews by Chakraborty et al. [12] and Hayes et al. [13] highlighted the importance of multi-omics integration of genomics, transcriptomics, epigenomics, proteomics, and metabolomics data for gaining a better understanding of tumor heterogeneity, biomarker identification, and personalized therapeutic modalities. Multi-omics models are always superior to single-omics models in terms of interpretability and robustness, especially when integrated with sophisticated machine learning algorithms.

However, the potential of multi-omics models is hindered by the presence of substantial computational complexities in terms of dimensionality, heterogeneity, and complexity of integration. Feature selection and efficient machine learning algorithms are required to make the most out of multi-omics data.

D. Research Gap and Positioning of the Proposed Work

From the existing literature, it is evident that while XGBoost-based models and deep learning approaches achieve high predictive accuracy, many studies either rely on single-omics data or lack robust external validation and interpretability.

The proposed work addresses these gaps by integrating multi-omics TCGA data with an XGBoost-driven feature selection framework, followed by hybrid classification using Logistic Regression and Support Vector Machines. Rigorous stratified cross-validation was used to ensure reliable performance estimation. Consequently, the proposed framework contributes toward robust, interpretable, and clinically relevant cancer classification in the context of precision oncology.

IV. System Architecture

The proposed system is an end-to-end machine learning–based cancer classification framework designed to distinguish between cancerous (Primary Tumour) and non-cancerous (Normal) samples using high-dimensional genomic expression data. The architecture follows a modular, reproducible pipeline comprising data ingestion, preprocessing, feature selection using Extreme Gradient Boosting (XGBoost), hybrid classification, stratified cross-validation, performance evaluation, and biomarker interpretation.

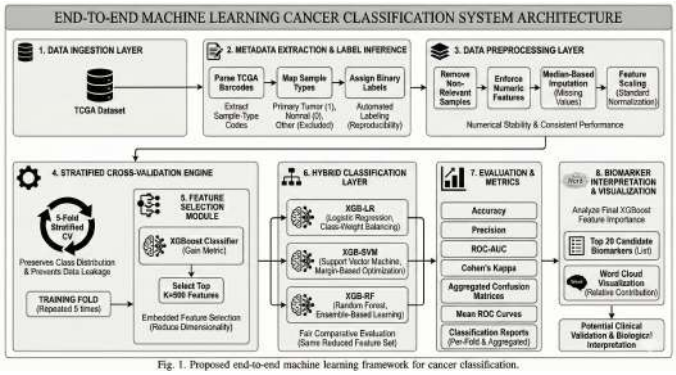

Fig. 1. Proposed end-to-end machine learning framework for cancer classification.

Fig. 1. System Architecture

To ensure a fair and unbiased comparison, three hybrid model variants are implemented and evaluated under identical experimental conditions: XGBoost combined with Logistic Regression (XGB–LR), XGBoost combined with Support Vector Machine (XGB–SVM), and XGBoost combined with Random Forest (XGB–RF).

1) Data Ingestion Layer: The input to the system is a gene expression matrix in CSV format obtained from The Cancer Genome Atlas (TCGA). Each row corresponds to a patient sample represented by a TCGA barcode, while columns correspond to gene expression features, comprising approximately 37,000 genes. This layer ensures correct sample–feature orientation (samples $\times$ features) and consistent string-based indexing to maintain compatibility with downstream machine learning libraries.

2) Metadata Extraction and Label Inference: TCGA barcodes are parsed to extract sample-type codes, which are subsequently mapped to biological categories. Samples are classified as *Primary_Tumour*, *Normal*, or *Other*, where non-relevant categories are excluded from further analysis. Binary labels are assigned such that Primary Tumour samples are encoded as 1 and Normal samples as 0. This automated labelling strategy eliminates manual annotation errors and enhances scalability and reproducibility.

3) Data Preprocessing Layer: The preprocessing stage prepares the data for reliable model training and convergence. Non-relevant samples are removed to restrict the task to binary classification. All gene expression features are enforced to be numeric, and missing values are imputed using median-based imputation to minimize the influence of outliers. Feature scaling is performed using standard normalization, which is essential for gradient-based and margin-based classifiers such as Logistic Regression and Support Vector Machines. These steps collectively ensure numerical stability and consistent model performance.

4) Stratified Cross-Validation Engine: To obtain robust and unbiased performance estimates, stratified five-fold cross-validation is employed. This strategy preserves class distribution across folds and prevents data leakage by performing feature selection strictly within each training fold. Such a design improves the generalization capability of the proposed framework and reduces optimistic bias in evaluation.

5) Feature Selection Module: Within each training fold, an XGBoost classifier is trained to compute feature importance scores using the Gain metric. Based on these scores, the top $K = 500$ most informative genes are selected. XGBoost is particularly well suited for this task due to its ability to efficiently handle high-dimensional data, capture non-linear gene–gene interactions, and perform embedded feature selection. This step substantially reduces dimensionality while retaining the most discriminative biological signals.

6) Hybrid Classification Layer: The reduced feature set obtained from the XGBoost module is used to train three independent classifiers. The XGB–LR model employs a linear decision boundary and offers high interpretability, with class-weight balancing to address data imbalance. The XGB–SVM model uses a margin-based optimization strategy that is effective in high-dimensional spaces and robust to overfitting. The XGB–RF model leverages ensemble-based learning to capture complex non-linear feature interactions while reducing variance through bagging. All classifiers operate on the same selected feature set, enabling a fair comparative evaluation.

7) Evaluation and Metrics: Model performance is assessed for each fold and classifier using accuracy, precision, area under the receiver operating characteristic curve (ROC–AUC), and Cohen's Kappa statistic. Cohen's Kappa is included to account for chance agreement, which is particularly important in imbalanced biomedical datasets. In addition, aggregated confusion matrices, mean ROC curves with pooled AUC, and per-fold as well as aggregated classification reports are generated to provide a comprehensive evaluation.

8) Biomarker Interpretation and Visualization: To enhance interpretability and translational relevance, feature importance scores from the final XGBoost model are analyzed. The top 20 genes are reported as candidate biomarkers, and word cloud visualizations are used to illustrate the relative contribution of highly ranked genes. These analyses support downstream biological interpretation and potential clinical validation.

V. METHODOLOGY

A. Study Design

This study presents an machine learning framework for cancer classification and biomarker identification using high-dimensional genomic data. The proposed methodology is inspired by the XGB-BIF framework introduced by Ghuriani et al. [9] and is extended to align with recent advances in multi-omics integration and cancer systems biology [12].

The primary objective of this work is twofold: (i) to develop a robust classification pipeline capable of distinguishing tumor from normal samples under severe class imbalance, and (ii) to identify biologically meaningful gene signatures that can serve as potential cancer biomarkers.

The overall workflow consists of the following sequential stages:

1) Data acquisition from publicly available cancer repositories
2) Data preprocessing and normalization
3) Feature selection using Extreme Gradient Boosting
4) Supervised classification using hybrid ensemble models
5) Model evaluation and cross-validation
6) Biomarker identification and biological interpretation
7) Extension toward multi-omics integration for enhanced robustness

All experiments were conducted using stratified cross-validation to ensure unbiased performance estimation under class imbalance.

B. Datasets

1) Gene Expression Datasets: Gene expression datasets corresponding to breast cancer and lung cancer were obtained from publicly available repositories and curated to include only samples annotated as either primary tumor or normal tissue.

The lung cancer gene expression dataset consists of 562 samples, including 511 primary tumor samples and 51 normal samples. Each sample is represented by 31,506 gene expression features.

The breast cancer gene expression dataset contains 1,224 samples, comprising 1,111 tumor samples and 113 normal samples. Each sample includes 31,575 gene expression features.

TABLE I
SUMMARY OF GENE EXPRESSION DATASETS

Cancer Type	Samples	Tumor	Normal	Features
Lung Cancer	562	511	51	31,506
Breast Cancer	1,224	1,111	113	31,575

2) Multi-Omics Datasets: To evaluate the robustness of the proposed framework beyond transcriptomic data alone, corresponding multi-omics datasets for both breast and lung cancer were obtained from publicly available repositories. These datasets contain heterogeneous molecular features from multiple omics source, capturing complementary biological signals across multiple omics layers.

The lung cancer multi-omics dataset retains the same 562 samples used in the gene expression analysis, while the breast cancer multi-omics dataset includes all 1,224 samples from the corresponding transcriptomic cohort. Although the integrated feature space is significantly higher-dimensional than gene expression alone, this representation enables the modeling of regulatory, genomic, and functional interactions that are not captured by transcriptomic profiles.

TABLE II
SUMMARY OF MULTI-OMICS DATASETS

Cancer Type	**Samples**	**Tumor**	**Normal**	**Features**
Lung Adenocarcinoma (LUAD)	592	533	59	37,600
Lung Squamous Cell Carcinoma (LUSC)	551	502	49	37,600
Breast Cancer (BRCA)	1,189	1,076	113	37,600

C. Preprocessing and Validation Strategy

All datasets underwent a standardized preprocessing pipeline prior to feature selection and model training. Non-numeric attributes were coerced into numeric format, and missing values were imputed using feature-wise median imputation to minimize bias introduced by incomplete measurements.

Given the severe class imbalance present in all datasets, models were evaluated using five-fold stratified cross-validation. This strategy preserves the original tumor-to-

normal ratio within each fold and enables reliable estimation of model generalization performance.

D. Feature Selection

1) XGBoost-Based Feature Ranking: Feature selection was performed using Extreme Gradient Boosting, a tree-based ensemble learning algorithm well suited for high-dimensional and nonlinear genomic data [9]. XGBoost ranks features using the gain importance metric, which measures the contribution of each feature to error reduction during tree construction.

Genes were ranked in descending order of importance, and the following top-ranked feature subsets were evaluated:

- Top 10 genes
- Top 50 genes
- Top 100 genes
- Top 500 genes
- Top 1000 genes

Empirical evaluation demonstrated that classification performance improved as the number of selected features increased but saturated beyond approximately 500 genes. Consequently, the top 500 XGBoost-ranked features were selected as the optimal subset for downstream classification.

2) Comparative Feature Selection Methods: To validate the effectiveness of XGBoost-based feature ranking, additional feature selection methods were implemented for comparison:

- Least Absolute Shrinkage and Selection Operator (LASSO), an embedded method based on L1 regularization
- Recursive Feature Elimination, a wrapper-based feature selection technique
- Variance threshold filtering to remove low-variance features

Comparative analysis revealed that XGBoost-based feature ranking consistently achieved superior classification accuracy and higher Cohen's Kappa values compared to LASSO, recursive feature elimination, and variance-based filtering.

E. Classification Models

Following feature selection, supervised machine learning classifiers were trained to distinguish between cancerous and non-cancerous samples. The following models were employed:

- Logistic Regression
- Random Forest
- Support Vector Machine

These models were selected due to their established effectiveness in gene-expression-based cancer classification tasks [1]. Hybrid ensemble combinations were evaluated by coupling XGBoost-based feature selection with each classifier:

- XGB + Logistic Regression
- XGB + Random Forest
- XGB + Support Vector Machine

Across both cancer types and data modalities, ensemble combinations involving Random Forest and Support Vector Machine demonstrated the most stable and robust performance.

F. Model Hyperparameters

To ensure reproducibility and transparency, the primary hyperparameters used in all experiments are summarized in Table III. Default parameters were retained where not explicitly specified.

TABLE III
HYPERPARAMETER CONFIGURATION OF CLASSIFICATION MODELS

Model	Hyperparameter	Value
XGBoost	n_estimators	300
	max_depth	6
	learning_rate	0.05
	subsample	0.8
	colsample_bytree	0.8
	random_state	42
Random Forest	n_estimators	500
	max_depth	None
	class_weight	balanced
	random_state	42
SVM	kernel	rbf
	C	1.0
	gamma	scale
	class_weight	balanced
Logistic Regression	penalty	l2
	solver	liblinear
	class_weight	balanced

G. Model Evaluation Metrics

Model performance was assessed using multiple evaluation metrics to capture different aspects of classification quality:

- Accuracy
- Area Under the Receiver Operating Characteristic Curve
- Cohen's Kappa coefficient

Cohen's Kappa was emphasized due to its robustness under class imbalance, as it measures agreement between predicted and true labels beyond chance.

H. Multi-Omics Integration Strategy

Although the primary analysis focused on transcriptomic data, the proposed framework aligns with established multi-omics integration principles described in recent literature [12]. Multi-omics integration involves combining complementary molecular data from genomics, transcriptomics, epigenomics, and proteomics to achieve a more comprehensive representation of tumor biology.

Integration strategies can be broadly categorized into vertical integration, horizontal integration, and diagonal integration. In this study, multi-omics datasets containing multiple molecular feature types were used as input to the classification models. Feature selection using XGBoost is employed to mitigate the challenges associated with high dimensionality and heterogeneous data sources.

This integration strategy enhances model robustness and biological interpretability, particularly for heterogeneous cancers such as breast cancer, and provides a foundation for future extensions toward clinically deployable precision oncology systems.

VI. EXPERIMENTAL RESULTS

All results are reported as mean ± standard deviation across five stratified cross-validation folds. Reporting variability across folds provides a more reliable estimate of model stability and generalization performance under class imbalance.

A. Single-Omics Classification Results

Gene expression–based models demonstrated strong discriminative performance across both cancer types. Table IV summarizes the mean classification metrics across classifiers.

TABLE IV
GENE EXPRESSION–ONLY CLASSIFICATION PERFORMANCE (MEAN ± STD)

Cancer	Model	Accuracy	Kappa
Lung	XGB + LR	0.9787 ± 0.0062	0.8851 ± 0.0214
Lung	XGB + RF	0.9947 ± 0.0038	0.9683 ± 0.0121
Lung	XGB + SVM	0.9964 ± 0.0029	0.9780 ± 0.0097
Breast	XGB + LR	0.9894 ± 0.0048	0.9402 ± 0.0156
Breast	XGB + RF	0.9927 ± 0.0041	0.9555 ± 0.0138
Breast	XGB + SVM	0.9943 ± 0.0035	0.9665 ± 0.0114

The XGB–RF model achieved an accuracy of 0.9927 ± 0.0041 and ROC–AUC of 0.9994 ± 0.0012 on the breast cancer dataset, demonstrating strong predictive performance and stability across folds. Similarly, non-linear classifiers (Random Forest and SVM) consistently outperformed Logistic Regression in terms of Cohen's Kappa, indicating improved agreement beyond chance in high-dimensional feature spaces.

1) Biomarker Analysis: XGBoost-based feature ranking identified biologically relevant genes. For lung cancer, dominant genes included *SUSD2*, *ANGPTL1*, *SHOX2*, *SFTPC*, and *CLDN18*, which are associated with epithelial differentiation, angiogenesis, and tumor progression.

For breast cancer, top-ranked biomarkers such as *MMP11*, *PAMR1*, *VEGFD*, *ADAMTS5*, and *CD300LG* are linked to extracellular matrix remodeling and vascular signaling pathways, supporting the biological plausibility of the selected features.

B. Multi-Omics Classification Results

Table V summarizes the mean performance metrics for integrated multi-omics datasets.

TABLE V
MULTI-OMICS CLASSIFICATION PERFORMANCE (MEAN ± STD)

Cancer	Model	Accuracy	Kappa
Lung	XGB + LR	0.9596 ± 0.0118	0.7770 ± 0.0324
Lung	XGB + RF	0.9835 ± 0.0065	0.8791 ± 0.0217
Lung	XGB + SVM	0.9706 ± 0.0097	0.7647 ± 0.0289
Breast	XGB + LR	0.9798 ± 0.0069	0.8937 ± 0.0194
Breast	XGB + RF	0.9933 ± 0.0041	0.9601 ± 0.0135
Breast	XGB + SVM	0.9823 ± 0.0062	0.8930 ± 0.0186

Classification using multi-omics datasets maintained high predictive performance while introducing modest variability due to increased feature heterogeneity. The XGB–RF model achieved an accuracy of 0.9933 ± 0.0041 and ROC–AUC of 0.9986 ± 0.0012 on the breast cancer multi-omics dataset, demonstrating both robustness and stability across folds.

1) Biomarker Analysis: XGBoost-based feature ranking applied to multi-omics data identified biologically relevant biomarkers across cancer types. The integration of multiple omics layers enables improved capture of molecular mechanisms underlying tumor progression. For breast cancer (BRCA), key biomarkers included *VEGFD*, *MMP11*, *PAMR1*, *CD300LG*, and *ADAMTS5*, which are associated with angiogenesis, extracellular matrix remodeling, and tumor invasion. These genes play important roles in vascular signaling and cancer progression. For lung adenocarcinoma (LUAD), dominant features such as *SFTPC*, *AGER*, *EZH2*, *FAM83A*, and *CAV1* are linked to epithelial differentiation, epigenetic regulation, and tumor proliferation. These biomarkers are well-established indicators of lung cancer biology. For lung squamous cell carcinoma (LUSC), important biomarkers included *SUSD2*, *ANGPTL1*, *CLDN18*, and *TEK*, which are associated with epithelial integrity, angiogenesis, and tumor micro-environment regulation.

VII. CONCLUSION

This study presented a reproducible machine learning framework for cancer classification and biomarker discovery using high-dimensional transcriptomic and multi-omics datasets. The proposed approach combines XGBoost-based embedded feature selection with supervised machine learning classifiers, including Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM), enabling substantial dimensionality reduction while preserving biologically informative features. The framework was evaluated on TCGA breast cancer (BRCA) and lung cancer cohorts, including lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), using stratified five-fold cross-validation to ensure reliable and unbiased performance estimation.

Experimental results demonstrate that transcriptomics-based classification provides strong discriminative capability for tumor versus normal sample identification. On the BRCA dataset, the XGB-RF model achieved a mean accuracy of 0.9933 ± 0.0041 and a Cohen's Kappa value of 0.9601 ± 0.0135, indicating high predictive stability and agreement beyond chance. Similarly, for lung cancer datasets, non-linear classifiers such as Random Forest and Support Vector Machine achieved accuracy values exceeding 0.98, with ROC-AUC values approaching 1.0, confirming the effectiveness of transcriptomic features for cancer classification.

Evaluation on multi-omics datasets also demonstrated consistently high predictive performance across classifiers. Although absolute accuracy improvements compared to transcriptomic data alone were modest in certain cases, multi-omics feature representations provided complementary molecular information and maintained robust classification performance. Furthermore, XGBoost-based feature ranking identified biologically relevant candidate biomarkers, including VEGFD, MMP11, PAMR1, AGER, EZH2, SFTPC, and AGRP, supporting the biological relevance and interpretability of the selected feature subsets.

Overall, the results indicate that XGBoost-based feature selection combined with ensemble and margin-based classifiers provides an effective, scalable, and interpretable approach for high-dimensional cancer classification. The proposed framework demonstrates strong predictive performance across multiple cancer types and data modalities, while also enabling biologically meaningful feature identification. Future work will focus on validating the framework using independent external datasets and exploring advanced multi-omics modelling approaches to further enhance generalization and clinical applicability.

VIII. Future Work

However, there are still some areas that need to be explored in the future.

First, it is important to validate the proposed framework on external datasets to evaluate the generalizability of the results. Validation on multiple datasets will help to strengthen the claim of its clinical utility.

Second, more sophisticated multi-omics integration approaches can be explored. Instead of using feature concatenation, more advanced integration techniques such as graph-based fusion learning, deep representation learning, and attention-based multi-modal neural networks can be employed to capture the interactions between different omics.

Third, biological pathway enrichment and network analysis can be extended to explore the relationships between the top-ranked genes. By integrating protein-protein interaction networks and regulatory pathway databases, the results may be more interpretable and provide potential therapeutic targets.

Fourth, survival analysis and prognostic modelling can be added to the framework to go beyond binary classification. Using Cox proportional hazards models or deep survival networks, the framework can be used for risk stratification and outcome prediction.

Fifth, SHAP (SHapley Additive exPlanations) analysis can be used to provide gene-level contribution scores to enhance clinical interpretability.

Finally, the prospective clinical validation and implementation within decision support systems would be the ultimate goal of translation. Implementation within digital pathology workflows and hospital information systems would help in the real-time classification of tumours and treatment strategies based on biomarkers.

In conclusion, this study provides a strong basis for the integrative classification of cancer based on machine learning and multi-omics data. Further improvements and validation would be required to translate the advances made within the computational field to the field of precision oncology.

References

[1] **Alharbi, F., & Vakanski, A.** (2023). Machine learning methods for cancer classification using gene expression data: A review. *Bioengineering*, 10(2), 173. https://doi.org/10.3390/bioengineering10020173

[2] **Sathishkumar, K., Chaturvedi, M., Das, P., Stephen, S., & Mathur, P.** (2022). Cancer incidence estimates for 2022 and projection for 2025: Result from National Cancer Registry Programme, India. *Indian Journal of Medical Research*, 156(4–5), 598–607. https://doi.org/10.4103/ijmr.ijmr_1821_22

[3] **Zhang, B., Shi, H., & Wang, H.** (2023). Machine learning and AI in cancer prognosis, prediction, and treatment selection: A critical approach. *Journal of Multidisciplinary Healthcare*, 16, 1779–1791. https://doi.org/10.2147/JMDH.S410301

[4] **Bhinder, B., Gilvary, C., Madhukar, N. S., & Elemento, O.** (2021). Artificial intelligence in cancer research and precision medicine. *Cancer Discovery*, 11(4), 900–915. https://doi.org/10.1158/2159-8290.CD-21-0090

[5] **Yu, C., & Helwig, E. J.** (2022). The role of AI technology in prediction, diagnosis and treatment of colorectal cancer. *Artificial Intelligence Review*, 55(1), 323–343. https://doi.org/10.1007/s10462-021-10034-y

[6] **Kumar, Y., Gupta, S., Singla, R., & Hu, Y.-C.** (2021). A systematic review of artificial intelligence techniques in cancer prediction and diagnosis. *Archives of Computational Methods in Engineering*. https://doi.org/10.1007/s11831-020-09483-y

[7] **McKinney, S. M., Sieniek, M., Godbole, V., et al.** (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6

[8] **Pantanowitz, L., Quiroga-Garza, G. M., Bien, L., et al.** (2020). An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images. *The Lancet Digital Health*, 2(8), e407–e416. https://doi.org/10.1016/S2589-7500(20)30159-X

[9] **Ghuriani, V., Wassan, J. T., Tripathi, P., & Chauhan, A.** (2025). XGB-BIF: An XGBoost-driven biomarker identification framework for detecting cancer using human genomic data. *International Journal of Molecular Sciences*, 26(12), 5590. https://doi.org/10.3390/ijms26125590

[10] **Zelli, V., Manno, A., Compagnoni, C., et al.** (2021). Classification of tumor types using XGBoost machine learning model. *Journal of Translational Medicine*, 19, 1–14. https://doi.org/10.1186/s12967-021-02844-6

[11] **Argelaguet, R., Marioni, J. C., Velten, B., et al.** (2018). Multi-omics factor analysis: A framework for unsupervised integration of multi-omics data sets. *Genome Biology*, 19, 25. https://doi.org/10.1186/s13059-018-1522-3

[12] **Chakraborty, S., Sharma, G., Karmakar, S., & Banerjee, S.** (2024). Multi-omics approaches in cancer biology: New era in cancer therapy. *Biochimica et Biophysica Acta*, 1870(5), 167120. https://doi.org/10.1016/j.bbadis.2024.167120

[13] **Hayes, C. N., Nakahara, H., Ono, A., Tsuge, M., & Oka, S.** (2024). From omics to multi-omics: A review of advantages and tradeoffs. *Genes*, 15(12), 1551. https://doi.org/10.3390/genes15121551

[14] **Hoadley, K. A., Yau, C., Hinoue, T., et al.** (2018). Cell-of-origin patterns dominate the molecular classification of tumors. *Cell*, 173(2), 291–304. https://doi.org/10.1016/j.cell.2018.03.022

[15] **Ramos, M., Schiffer, L., Re, A., et al.** (2017). Software for integration of multi-omics experiments in Bioconductor. *Cancer Research*, 77(21), e39–e42. https://doi.org/10.1158/0008-5472.CAN-17-0344

[16] **Molania, R.** (2022). TCGA PanCancer RNAseq dataset. *Zenodo*. https://doi.org/10.1101/2021.11.01.466731

[17] **UCSC Xena Consortium** (2023). TCGA Pan-Cancer Atlas Data Hub. *UCSC Xena Browser*. https://xenabrowser.net/

[18] **Liaw, A., & Wiener, M.** (2002). Classification and regression by randomForest. *R News*, 2(3), 18–22.

[19] **Singh, A., Shannon, C. P., Gautier, B., et al.** (2019). DIABLO: Integrative approach for multi-omics biomarker discovery. *Bioinformatics*, 35(17), 3055–3062. https://doi.org/10.1093/bioinformatics/bty1054

[20] **Cava, C., Salvatore, C., & Castiglioni, I.** (2023). Pan-cancer classification using artificial neural networks. *Applied Sciences*, 13, 7355. https://doi.org/10.3390/app13137355

[21] **Dikaios, N.** (2022). Sparse-input neural networks to classify cancer types. *Onco*, 2, 56–68. https://doi.org/10.3390/onco2020005

[22] **Li, Y., Kang, Z., Zhang, X., & Li, J.** (2018). Cancer classification using genetic algorithm and k-nearest neighbor method based on gene expression data. *Computational Biology and Chemistry*, 74, 248–255. https://doi.org/10.1016/j.compbiolchem.2018.03.014

[23] **Wang, J., Dai, X., Luo, H., Yan, C., Zhang, G., & Luo, J.** (2021). MI DenseNetCAM: A novel pan-cancer classification and prediction method based on mutual information and

deep learning model. *Computational Biology and Chemistry*, 92. https://doi.org/10.1016/j.compbiolchem.2021.107457

[24] **Lyu, X., Li, Y., & Feng, J.** (2022). Deep learning-based classification of cancer types using gene expression data. *Frontiers in Genetics*, 13. https://doi.org/10.3389/fgene.2022.856764

[25] **Chen, T., & Guestrin, C.** (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 785–794). https://doi.org/10.1145/2939672.2939785