Journal of Genetic Engineering and Biotechnology 24 (2026) 100639

Contents lists available at ScienceDirect

Journal of Genetic Engineering and Biotechnology

journal homepage: www.elsevier.com/locate/jgeb

AI based multiomics integration for cancer diagnosis and prognosis

Moshira Ghaleb *, Maryam Al-Berry, Hala Ebied, Mohamed Tolba

Scientific Computing Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt

ARTICLE INFO

Keywords:
Multiomics
Pancancer
Ovarian Cancer
Deep Learning
Machine Learning
OmicsFusionNet
AI

ABSTRACT

Cancer remains a critical global health challenge, driving the need for innovative approaches in diagnosis and treatment. This research introduces OmicsFusionNet, an AI-powered hybrid model integrating machine learning and deep learning to revolutionize cancer care. The tool incorporates up to six multiomics datasets—genomics, transcriptomics, and epigenomics—achieving 80.2% accuracy across 23 cancer types. Notably, RNAseq and methylation integration reached 99.8% accuracy, highlighting XGBoost's feature selection and deep learning's classification strength.

For ovarian cancer stage detection, OmicsFusionNet optimized analysis using CPTAC-OV and TCGA-OV datasets, achieving accuracies between 83% and 91% by combining ElasticNet and XGBoost with deep learning. Additionally, KEGG pathway enrichment of multiomics biomarkers identified key cancer-related pathways, advancing early detection, biomarker discovery, and personalized treatments.

This study underscores the transformative potential of AI and multiomics integration in cancer research, enabling precise interventions and uncovering novel mechanisms that enhance patient outcomes.

1. Background

The global cancer burden continues to escalate, with lung, breast, and colorectal cancers remaining the most prevalent worldwide.[1] According to the WHO, cancer persists as a primary cause of mortality globally, with breast cancer being the most common malignancy among women, while lung cancer holds the highest fatality rate in men. Pediatric cancers, such as lymphomas, brain tumors, and leukemia, also pose significant risks. Ovarian cancer, ranking eighth in both prevalence and mortality among women, is particularly challenging due to its frequent late-stage diagnosis.[2] Accurate classification of cancer types is pivotal for determining appropriate treatment strategies and improving survival outcomes.[3]

With the rising incidence and mortality rates, the demand for advanced diagnostic techniques capable of accurately identifying cancer types and stages has become evident. Multiomics data analysis,[4] in conjunction with machine learning (ML)[5] and deep learning (DL),[6] offers the potential to significantly enhance the precision and effectiveness of cancer diagnostics, thereby improving patient management and treatment outcomes.[7] These technologies have revolutionized imaging, data processing, and biological analysis by automating image interpretation and enhancing the detection of anomalies, leading to more accurate insights from medical imaging modalities such as MRI, CT scans, and X-rays.[8] As a result, early cancer detection and improved patient outcomes have been realized.[9,10]

In the field of data science, ML and DL methodologies are utilized to extract patterns, enabling real-time decision-making, personalized recommendations, and predictive analytics. Additionally, they are instrumental in identifying biomarkers within genomic, proteomic, and metabolomic datasets, fostering a more profound understanding of intricate biological processes.[11,12] These methods not only expedite drug discovery but also predict disease progression and support precision medicine by tailoring treatments to the unique genetic profiles of individual patients [13,14].

The integration of omics data[15] is crucial for comprehending complex diseases such as cancer. Omics encompasses genomics, transcriptomics, proteomics, metabolomics, and epigenomics, with DNA methylation[16] playing a key role in gene regulation. When methylation data is combined with other omics datasets, insights into disease progression are obtained that single-omics approaches cannot provide. Various omics integration strategies have been employed, often leveraging ML techniques, including early, intermediate, and late integration methods. These approaches enhance findings across different omics layers, allowing for stronger conclusions with the application of

* Corresponding author at: Scientific Computing Department, Faculty of Computer and Information, Science, Ain Shams university, Nasr city, Cairo, Egypt.
E-mail addresses: moshirasg@cis.asu.edu.eg (M. Ghaleb), maryam_nabil@cis.asu.edu.eg (M. Al-Berry), Halam@cis.asu.edu.eg (H. Ebied), fahmytolba@cis.asu.edu.eg (M. Tolba).

https://doi.org/10.1016/j.jgeb.2025.100639
Received 29 November 2024; Received in revised form 21 July 2025; Accepted 11 December 2025

ML algorithms.[17]

The use of ML techniques improves the accuracy of cancer classification by integrating multiomics data, enabling more comprehensive analysis and superior predictive performance. For instance, Garcia-Diaz et al.[18] applied a genetic algorithm for feature selection across five cancers using RNAseq data, achieving an average accuracy of 98.81%. Similarly, Ramaswamy et al.[19] employed support vector machines (SVMs) with recursive feature elimination for multiclass cancer classification, resulting in 78% accuracy. Lee et al.[20] developed an ensemble classifier, CPEM, which attained 84% accuracy across 31 cancer types using TCGA data.

Tabares-Soto et al.[21] compared ML and DL methods for the classification of 11 tumor types, revealing that DL outperformed traditional ML approaches. Zeng et al.[22] achieved 77.6% accuracy in classifying seven cancers using convolutional neural networks (CNNs). In contrast, logistic penalized linear regression and linear SVM methods demonstrated lower accuracy, underscoring the superiority of advanced techniques such as DL. Castillo et al.[23] further emphasized the importance of feature selection, utilizing RNAseq data to classify leukemia effectively. Ghaleb et al.[24] applied integrated proteomic datasets for ovarian cancer prediction, achieving 99% accuracy when combining ElasticNet with SVM.

Xie et al.[25] used RNAseq and proteome data to identify ovarian cancer subtypes, while Bismeijer et al.[26] introduced a modified iCluster model to identify key processes within tumor collections through multiomics integration. Multiclustering methods like NEMO[27] and MONET[28] further advanced clustering approaches in cancer datasets, revealing distinct patterns in multiomics data. Jeong et al.[29] used RNA sequencing and clinical data for ovarian cancer stage prediction, achieving AUC scores between 0.700 and 0.813, while Munetoshi et al.[30] demonstrated the superiority of XGBoost in ovarian cancer prognosis prediction, achieving an accuracy of 0.809. Despite promising results in the literature, optimizing minimal omics data use while maintaining high accuracy across various cancer types is still needed for clinical feasibility.

In this study, an AI-driven multiomics tool for pancancer classification[31] is presented, capable of accurately identifying various cancer types, including ovarian, breast, and lung cancers. To demonstrate the tool's practical utility beyond classification, it was further applied to predict the stage of ovarian cancer patients, with ovarian cancer chosen as a representative case study. The task of ovarian cancer stage detection was used to showcase how the tool can support downstream clinical decision-making after cancer type identification. Ovarian cancer was selected for stage detection due to its high clinical significance, the availability of comprehensive public multi-omics datasets, and its well-documented molecular heterogeneity, making it an ideal representative for evaluating the applicability and robustness of the proposed tool in a real-world context. The model integrates multiple ML models for feature selection from individual omics, with the selected features combined for further analysis. A DL model is then employed for classification, and its accuracy is compared with the Support Vector Machine (SVM),[32] demonstrating the model's efficiency in handling complex datasets. The remainder of this paper is organized as follows: Section Two outlines the methodology, including the OmicsFusionNet proposed models; Section Three presents' results and discussion; Section Four concludes the study, and section five present the limitation and the future research directions.

2. Methodology

The main purpose of this study was to develop a robust intelligent model for detecting cancer types when applied to various types of cancer, as well as for identifying cancer stages when applied to a single cancer type. The study also aimed to highlight important genes as biomarkers for specific cancer types and to utilize selected multi-omics datasets to achieve higher detection accuracy. OmicsFusionNet, a hybrid model that integrates ML and DL techniques, was proposed to achieve high accuracy through the integration of various omics data.

OmicsFusionNet is composed of three key terms. "Omics" refers to the input of one or more types of omics data. "Fusion" describes the integration of ML for feature selection and DL for classification within a neural network framework. "Net" indicates the network architecture that enables this integration. The model was constructed and applied to various datasets for evaluation. Data preprocessing, feature selection, classification, and performance assessment were conducted. The model was designed to accommodate both single-omic and multi-omics data, with an emphasis on late-stage data integration.

OmicsFusionNet is composed of two primary components:

- Omics Data Integration: Up to six omics datasets (e.g., RNA-seq, methylation, Proteome) were integrated for analysis.
- Hybrid Models: integrate ML and DL algorithms, such as XGBoost[33] and ElasticNet,[34] were

employed for feature selection, while DL and SVM were applied for classification tasks depending on the dataset and classification problem (cancer type or stage). A Workflow Diagram of the Comprehensive Study on Multiomics Data Integration for Cancer Classification and Stage Prediction in Fig. 1.

This figure illustrates the comprehensive workflow, including multiomics data integration, feature selection, classification, prediction, and performance for the proposed models.

Multiple omics datasets were integrated to enhance feature extraction and data fusion, with varying levels of omics integration—three, four, five, and six-omics—applied based on the data available for each task.

- **Three-omics fusion**: RNA-seq, proteomics, and phosphoproteome datasets were merged.
- **Four-omics fusion**: RNA-seq, methylation, SCAN, and mutation data were combined.
- **Five-omics fusion**: RNA-seq, proteomics, methylation, SCAN, and mutation data were integrated.
- **Six-omics fusion**: RNA-seq, proteomics, methylation, SCAN, mutation, and phosphoproteome datasets were utilized.

This multi-layered integration enabled a more comprehensive understanding of cancer by capturing genetic, epigenetic, and proteomic insights, allowing for a deeper exploration of cancer mechanisms.

2.1. Preprocessing stage

Identical preprocessing steps were applied across all experiments to ensure consistency:

- Data Cleaning: Nonnumeric data were converted to numeric values, features with zero variance were filtered out, and missingness rates were computed.
- Missing Data Imputation: the missing data[35] were calculated, then apply the K-Nearest Neighbors (KNN) algorithm was used to impute missing values.[36]
- Normalization: Log transformation was applied to normalize the data.[37]
- Batch effect: All multi-omics datasets were obtained from The Cancer Genome Atlas (TCGA), where standardized protocols and centralized processing pipelines were applied across all omics platforms. Through this consistent approach, technical variability and batch effects across different omics types were minimized.[38] To further assess potential batch effects and to verify that the observed variation was primarily driven by biological factors, principal component analysis (PCA) was performed on the integrated multi-omics dataset, and the samples were visualized according to their respective cancer

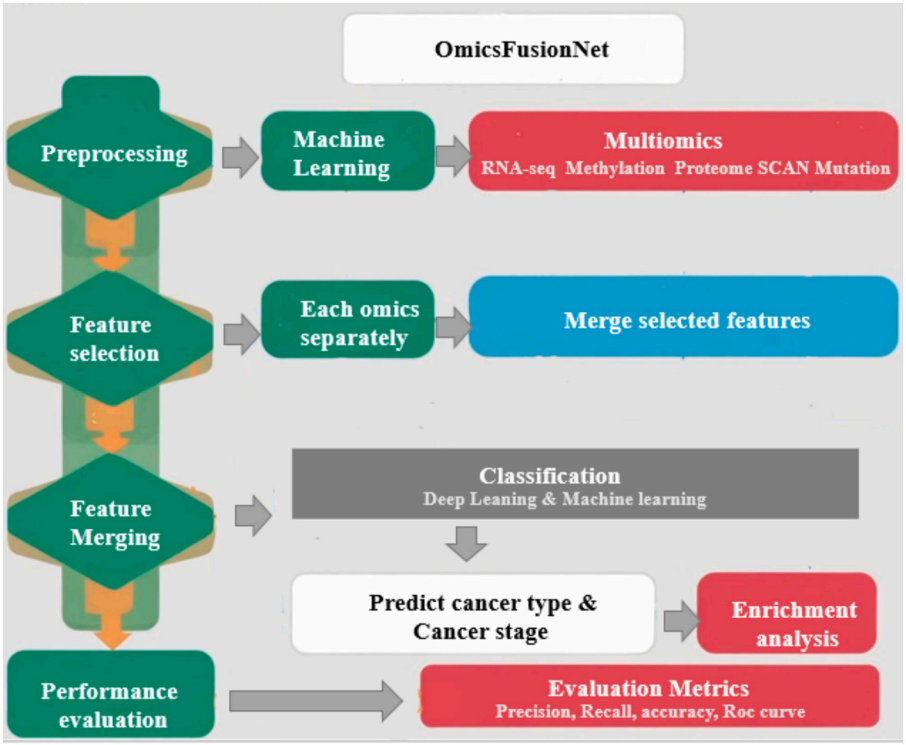


Fig. 1. Workflow Diagram of the Comprehensive Study on Multiomics Data Integration for Cancer Classification and Stage Prediction based on Omics Data Integration.

types. Fig. 2 PCA of the integrated multi-omics dataset, where samples are visualized based on their respective cancer types.

For the ovarian cancer stage detection task, two independent datasets from TCGA and the Clinical Proteomic Tumor Analysis Consortium (CPTAC) were utilized. To avoid potential data integration bias, the datasets were analyzed separately under distinct experimental settings. No direct merging of the datasets was conducted.

2.2. Feature selection phase

Three different techniques were applied during the feature selection phase. Each technique was
applied on single omic datasets.

2.2.1. *Xgboost (eXtreme Gradient Boosting)*

XGBoost is a ML algorithm that is considered highly effective for large datasets with numerous features, as it automatically ranks features by importance, facilitating the identification of the most relevant ones. Overfitting is effectively controlled through built-in L1 and L2 regularization, enhancing the model's robustness. High classification accuracy is achieved by iteratively improving weak learners and focusing on errors from previous models. Additionally, parallelization is utilized to significantly accelerate the training process, making it particularly advantageous for handling large datasets. XGBoost is also capable of handling unbalanced data. The default parameters were applied without any tuning.

$$\text{Obj(t)} = \sum_{i=1}^{1} L\left(y_i, y_i^{**t-1)} + f_t(x_i)\right) + \sum_{t=1}^{T} \Omega(f_t) \quad (1)$$

Where:

- L is the loss function, Mean squares.
  - $y_i^{**t-1)}$ is the predicted value from the previous iteration.
  - f_t is the new tree added in iteration t.
  - $\Omega(f_t)$: is the regularization term to penalize the complexity of the model.

2.2.2. *ElasticNet*

ElasticNet is a regularization technique commonly used in ML for linear regression models. It combines L1 (Lasso) and L2 (Ridge) penalties, making it effective for both feature selection and accuracy improvement. A balance is struck between shrinking coefficients and eliminating irrelevant features, with unimportant ones often reduced to zero for easier interpretation. Correlated features are handled more effectively than by Lasso or Ridge alone, and ElasticNet is particularly useful for small datasets with many features, helping to prevent overfitting. The default parameters were applied, with alpha tuned to 0.2 and l1_ratio to 0.005.

$$min_{\beta=\left(\frac{1}{2n}\sum_{i=1}^{n}\left(y_i - x_i^T\beta\right)^2 + \alpha((1-\lambda)\|\beta_1\| + \lambda\|\beta\|_2^2\right.} \quad (2)$$

Where:

- :y_i The true output for the i-th sample.
- x_i^T: The feature vector for the i-th sample.
- β: The vector of coefficients to be learned.
- $\|\beta_1\|$: The L1 norm (sum of the absolute values of the coefficients).
- $\|\beta\|_2^2$: The L2 norm (sum of the squared values of the coefficients).
- α: Controls the overall strength of regularization.
- λ: The mixing ratio between L1 and L2 regularization (also referred to as l1_ratio in some implementations).

2.2.3. *Merging XGBoost and ElasticNet*

During the feature selection stage offered multiple advantages due to the complementary strengths of each approach. Feature selection that captures both non-linear XGBoost and linear ElasticNet relationships is

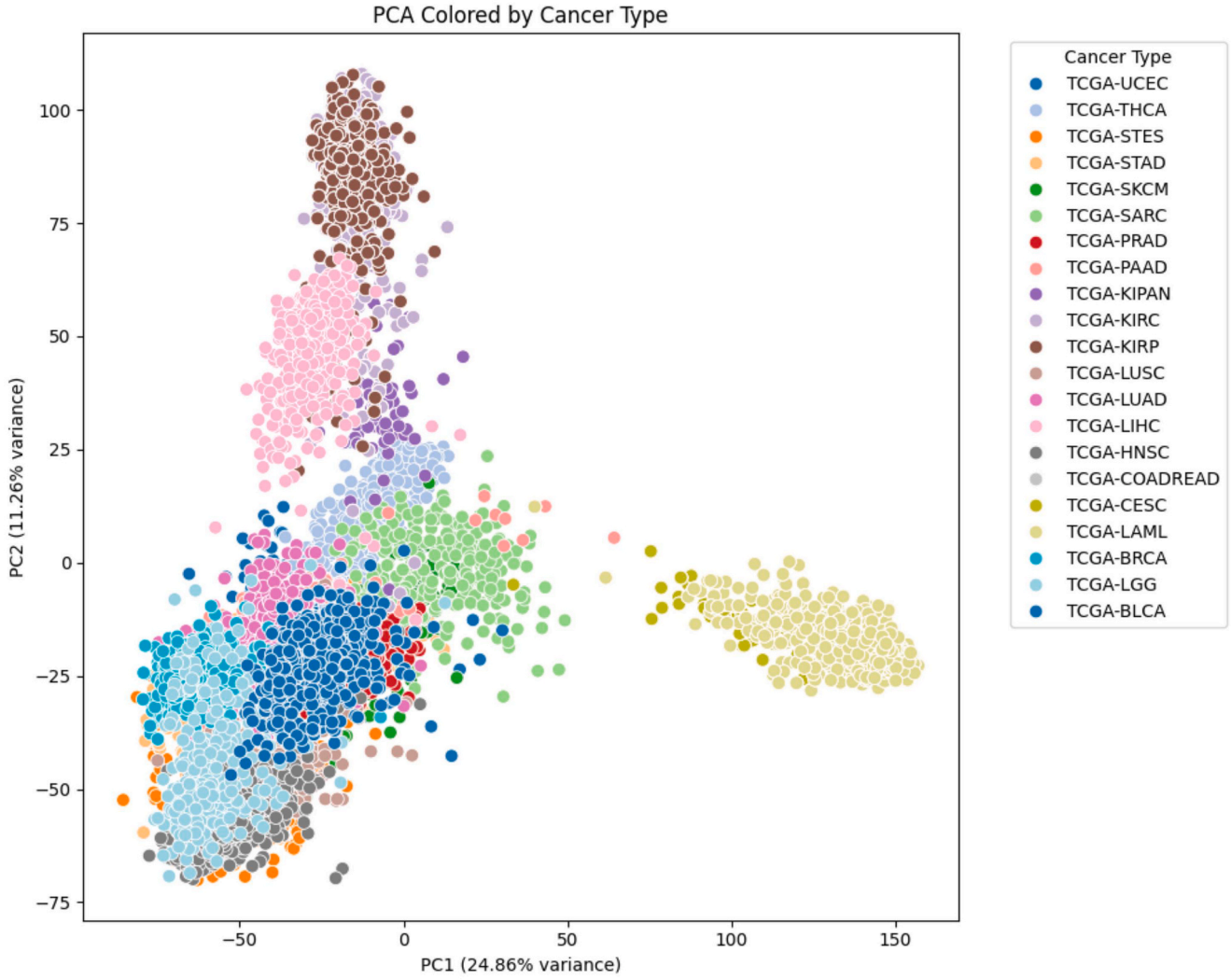


Fig. 2. Principal Component Analysis (PCA) of Integrated Multi-Omics Data Colored by Cancer Type.

ensured, resulting in a more robust process. A more interpretable and diverse feature set is created, with ElasticNet addressing correlations and XGBoost capturing complex interactions. Generalization is enhanced, as underfitting is prevented by XGBoost and overfitting is reduced by ElasticNet. Confidence in feature relevance is increased when both models select the same features, leading to a well-rounded and reliable selection process. Feature extraction from each omic is performed separately with ElasticNet and XGBoost, after which the top and bottom features selected by ElasticNet are merged with those selected by XGBoost to form a comprehensive feature set.

$$Features_{final} = Top_x(\beta_{ElasticNet}) \cup Bottom_x(\beta_{ElasticNet}) \cup \beta_{XGBoost} \quad (3)$$

Where:

- $Top_x(\beta_{ElasticNet})$: The top x features selected by ElasticNet based on feature importance.
- $Bottom_x(\beta_{ElasticNet})$: The bottom x features selected by ElasticNet.
- $\beta_{XGBoost}$: The all features selected by XGBoost.
- $\cup$ denotes the union of the feature sets.

2.3. Classification phase

Both SVM and DL were used in the models to compare classification performance. SVM was applied in three models, recognized for its effectiveness in handling high-dimensional data and its ability to perform well when a clear margin of separation exists, making it especially reliable for smaller datasets. In four models, DL was employed, known for excelling at learning complex patterns from large datasets and effectively managing tasks involving intricate feature interactions and non-linear relationships, particularly in large-scale data. Table 1 provides a detailed overview of the key architectural components for each DL model.

2.4. The full Architectures of the proposed OmicsFusionNet models

Six distinct models are incorporated within OmicsFusionNet, each utilizing different feature selection and classification algorithms. The

Table 1
The proposed deep learning networks for OmicsFusionNet models.

Convolutional Layers	Fully Connected Layers	Activation Functions	Output Layer	Application
5 layers with 256, 128, 64, 32, 16 filters	4 layers with 2048, 1024, 500, 23 unit	ReLU activations	SoftMax	Pancancer
3 layers with 128, 64, 32 filters	3 layers with 1024, 500, 1 unit	ReLU activations	Sigmoid	Ovarian Stage

model architecture is divided into three stages: preprocessing, feature selection, and classification. The same preprocessing steps were applied uniformly to all models and data. These models were applied to two tasks: pancancer classification and ovarian cancer stage classification. The first model addressed both tasks, while all six models were focused on ovarian cancer stage classification to compare performance. The details of these models are presented in Table 2.

The full architecture of Model 1, which is exclusively for pancancer classification, is shown in Fig. 3. For ovarian cancer stage classification, six models were developed; however, Model 6 is presented in Fig. 4 as it is the most complex of the six models.

The deep learning component of OmicsFusionNet was designed to integrate multi-omics data and to capture complex, nonlinear relationships across different molecular layers. The architecture is composed of a multi-branch neural network, in which each omics type (e.g., RNA-seq, methylation, mutation, SCAN) is processed through a dedicated input branch. Within each branch, a combination of convolutional layers and fully connected layers was utilized to extract both local and global feature patterns. These layers were followed by ReLU (Rectified Linear Unit) activation functions to introduce non-linearity. To reduce the risk of overfitting, dropout layers with rates ranging from 0.3 to 0.5 were applied after key layers. The final classification layer was implemented with a softmax activation function for multi-class classification tasks (pan-cancer classification), or a sigmoid activation function for binary classification tasks (e.g., ovarian cancer stage detection). The model was trained using the Adam optimizer with an initial learning rate of 0.001. Categorical cross-entropy was used as the loss function for multi-class tasks, while **binary cross-entropy** was used for binary tasks. To prevent over-training, early stopping was applied based on validation loss. Hyperparameters, including the number of layers, number of neurons per layer, and dropout rates, were tuned through grid search and cross-validation.

3. Experimental results and discussion

The datasets used in the study were processed and analyzed to evaluate model performance. Performance measurements, including accuracy, precision, recall, and F1 score, were calculated. The results are discussed in terms of their implications for the effectiveness of the proposed model.

3.1. Datasets

In this study, datasets comprising RNA sequencing (RNA-seq) and methylation data for various cancer types were sourced from The Cancer Genome Atlas (TCGA).[39] The research focused on 23 cancer types, utilizing two omics data types: RNA-seq and methylation. The RNA-seq dataset contained 10,042 samples, each with approximately 23,000 features, while the methylation dataset included 9,345 samples, each with around 20,000 features. For ovarian cancer stage classification, the CPTAC and TCGA datasets were employed, both obtained from Linked Omics.[40] The ovarian CPTAC dataset encompassed three omics types RNAseq, phosphoproteome, and proteome; comprising 75 samples for each omics type. The TCGA ovarian cancer dataset included six omics types: methylation, mutation, copy number variation (SCAN), proteome, RNA-seq, and phosphoproteome. Specifically, the TCGA dataset consisted of 364 samples for methylation, 302 for mutation, 361 for SCAN, 192 for RNA-seq, 42 for phosphoproteome, and 77 for proteome. This extensive dataset provides a robust foundation for multiomics analysis in ovarian cancer research.

Table 2
Summary of Feature Selection and Classification Approaches for Pancancer and Ovarian Cancer Stage Classification.

Model	No. omics used per experiment	Feature Selection	Classifier	Application
One	One or two	XGBoost	DL	Pancancer classification
One	Three or four or five or six	XGBoost	DL	Ovarian cancer stage classification
Two	Three or four or five or six	ElasticNet	SVM	Ovarian cancer stage classification
Three	Three or four or five or six	ElasticNet	DL	Ovarian cancer stage classification
Four	Three or four or five or six	XGBoost	SVM	Ovarian cancer stage classification
Five	Three or four or five or six	ElasticNet + XGBoost	SVM	Ovarian cancer stage classification
Six	Three or four or five or six	ElasticNet + XGBoost	DL	Ovarian cancer stage classification

3.2. Performance Criteria

The performance evaluation phase utilized an accuracy measure for all the experiments, and for the best experimental results, additional metrics included precision, as in (4); recall, as in (5); the F1 score, as in (6); and accuracy, as in (7).

$$\text{Precision} = \text{true positives} / \text{true positives} + \text{false positives} \tag{4}$$

$$\text{Recall (sensitivity)} = \text{true positives} / \text{true positives} + \text{false negatives} \tag{5}$$

$$\text{F1 Score} = 2 \times (\text{precision} \times \text{recal} / \text{precision} + \text{recal}) \tag{6}$$

$$\text{Accuracy} = (\text{the number of correct predictions} / \text{total number of predictions}) \times 100 \tag{7}$$

The ROC (Receiver Operating Characteristic) curve is used to evaluate the performance of a classification model by plotting the True Positive Rate (TPR) as in (8) against the False Positive Rate (FPR) as in (9) across different threshold settings. The model's ability to differentiate between two classes, such as cancer stages (e.g., Stage 3 versus Stage 4), is quantified by the AUC (Area Under the Curve). A higher AUC score signifies greater accuracy in distinguishing between these stages. True Positive Rate (TPR) (Sensitivity or Recall): It measures how many actual positives (e.g., Stage 4) are correctly identified by the model.

$$\text{TPR} = \text{True positive (TP)} / \text{True Positives (TP)} + \text{False Negatives (FN)} \tag{8}$$

$$\text{FPR} = \text{False Positive (FP)} / \text{False Positives (FP)} + \text{True Negatives (TN)} \tag{9}$$

The ROC curve provides a visual representation of how effectively the model differentiates between the two cancer stages, while the AUC score quantifies this performance. A score closer to 1 signifies a stronger ability to classify the stages accurately.

3.3. Pancancer classification results

The PCA plot of the integrated multi-omics dataset Fig. 2 was found to exhibit clear clustering of samples based on their biological cancer types (e.g., TCGA-BRCA, TCGA-LUSC, etc.). Samples corresponding to distinct cancer types were generally grouped into separate clusters, while some degree of overlap was observed among biologically similar cancers, as is expected in complex multi-omics data. Notably, no unexpected clustering patterns indicative of technical batch effects were detected. These findings indicate that the primary sources of variation within the dataset are biological in nature, thereby confirming its suitability for downstream pan-cancer classification tasks.

For Pancancer classification, the OmicsFusionNet Model 1 was utilized across three experiments, each addressing different data types and balances. The data were split into 80 % for training and 20 % for testing

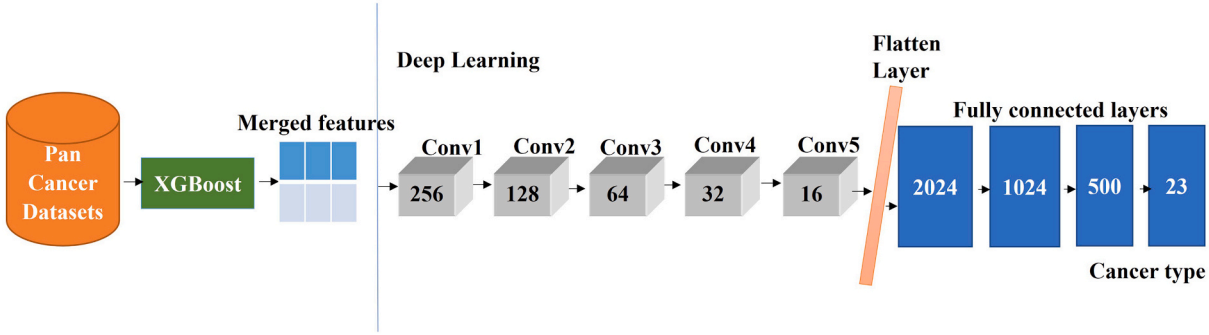


Fig. 3. The OmicsFusionNet Model One architecture for Pancancer classification.

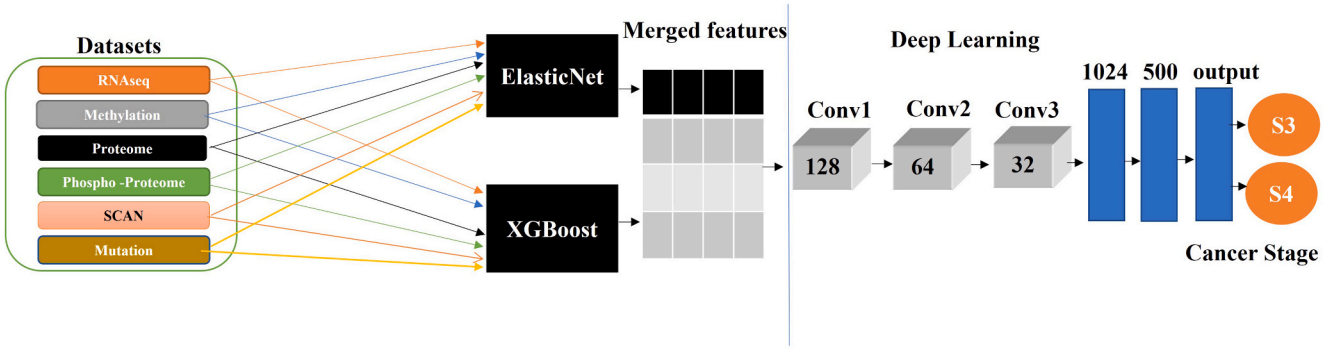


Fig. 4. The OmicsFusionNet Model six architecture for ovarian cancer stage classification.

in all experiments. Following consistent preprocessing steps across all experiments, feature selection and classification were performed. In the first experiment, 23 cancer types were analyzed using RNA-seq data with 10,042 samples. Despite unbalanced data, feature selection using XGBoost achieved 65.65 % accuracy, with 5,857 selected features. Deep learning classification followed, improving accuracy to 80.20 %. The second experiment addressed data imbalance by selecting 200 samples per cancer type, yielding 4,600 samples. The XGBoost model achieved an accuracy of 83.90 %, selecting 3,547 features. Applying deep learning led to a classification accuracy of 90.60 %, highlighting the importance of balancing datasets for model reliability.

These results emphasize the value of combining RNA-seq and methylation data for superior accuracy in cancer classification. A detailed comparison of feature selection and classification accuracy across the three experiments is presented in Table 3. The integration of Methylation and RNAseq data in cancer classification is offered as a robust approach to enhance predictive accuracy and model performance. By combining these two complementary data types, a more comprehensive biological landscape can be captured, which is crucial for understanding the complexities of cancer. This integration not only leads to improved feature extraction, achieving an accuracy of 99.80 %, but also results in a significant boost in classification performance across various cancer types. More precise identification of cancer subtypes is enabled by this synergy, and the discovery of novel biomarkers is facilitated.

3.3.1. *Performance analysis for OmicsFusionNet proposed models*

To further illustrate the effectiveness of this integrated approach, the precision and recall metrics for the 22 cancer classes are compared in Fig. 5, highlighting the model's ability to maintain a balance between false positives and false negatives. Additionally, Fig. 6 presents a bar graph showcasing the F1 score for the same 22 classes, providing a consolidated view of the model's performance across different cancer types. The advantages of using integrated Methylation and RNAseq data for enhanced cancer classification outcomes are underscored by these visualizations. Additionally, the selected features from the ML and DL algorithms, which contribute to the high classification accuracy, could serve as novel biomarkers for the 22 cancer types.

In our study on pancancer classification across 22 cancer types, we depict the model's performance in the confusion matrix Fig. 7. This matrix presents the counts of true and predicted classifications, with each row representing the actual cancer type and each column indicating the model's predictions.

The strong diagonal pattern in the matrix highlights the model's accuracy in correctly identifying several cancer types. High values along the diagonal, seen for cancer types like STES (196), KIPAN (239), and BLCA (205), indicate successful predictions where the true and predicted classes align. Off-diagonal entries reveal instances of misclassification, such as occasional misclassifications of KIPAN and LUSC, which suggests areas where further refinement or additional data could enhance class separability. Notably, our results indicate that the selected

Table 3
Summary of Experiments for Pancancer Classification.

Exp.	Data Type	No. Samples	Features Identified	Feature Selection Model	Selected Features	XGBoost Accuracy	DL Accuracy	No. Class
1st	RNA-seq	10,042	18,983	XGBoost	5,857	65.65 %	80.20 %	23
2nd	RNAseq	4,600	18,983	XGBoost	3,547	83.90 %	90.60 %	23
3rd	RNA-seq + Meth	9,345	20,079 (Meth), 18,983 (RNA-seq)	XGBoost	402 (Meth), 5,857 (RNA-seq	100 % (Meth), 83.90 % (RNAseq)	99.80 %	22

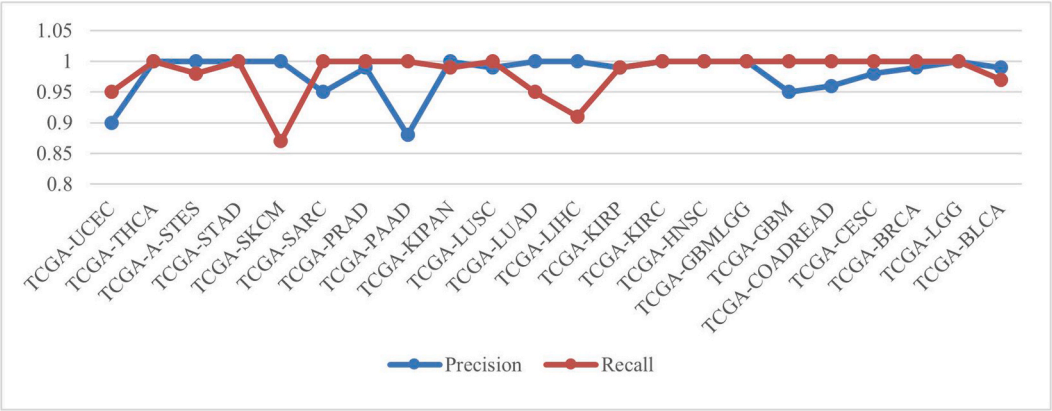


Fig. 5. Precision and recall of 22 Cancer Types with Methylation and RNA Data Integration.

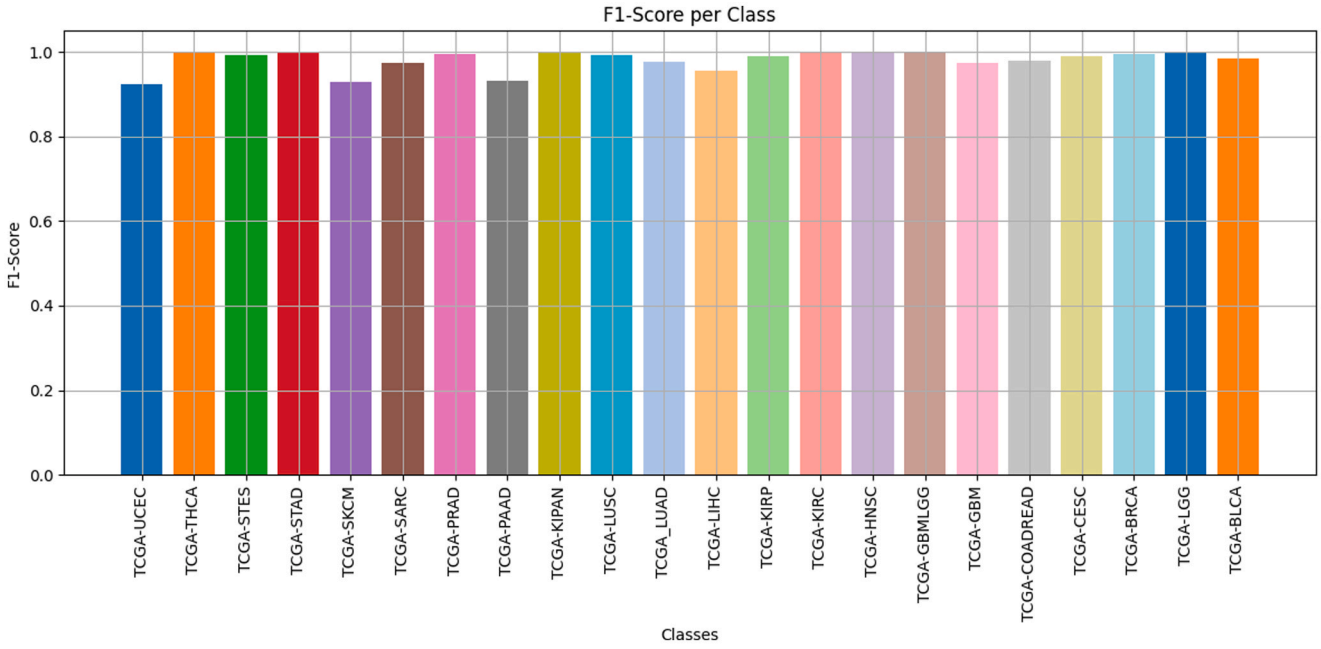


Fig. 6. F1 Score Comparison Across 22 Cancer Types from Methylation and RNA Integrated Model.

features in this classification model serve as reliable biomarkers for these cancer types. These biomarkers enhance early detection, a crucial factor for improving patient outcomes. Additionally, the model's capacity to distinguish between various cancer types underscores the potential value of these biomarkers in guiding treatment decisions, supporting more personalized and effective therapeutic approaches. Overall, this analysis of the confusion matrix demonstrates the model's effectiveness in handling multi-class classification within a complex pancancer setting, while also identifying areas for future enhancement. The results emphasize both the model's strengths and the promising role of identified biomarkers in early detection and treatment stages.

To further assess the classification performance, ROC curves were generated for each cancer type in the pan-cancer classification task. As shown in Fig. 8, an AUC of 1.00 was achieved across all 22 cancer classes, indicating that perfect discrimination between cancer types was obtained. This outcome demonstrates that high true positive rates with minimal false positives were consistently achieved. The results confirm that strong predictive performance and generalization were attained by the proposed model, thereby validating the effectiveness of the integrated multi-omics framework and the robustness of the employed classification strategy.

The results of our experiments demonstrate the substantial benefits of a hybrid multi-omics approach for pancancer classification. By leveraging RNA-seq and methylation data, OmicsFusionNet showed notable improvements in classification accuracy across three experiments, underscoring the value of integrating diverse data types to capture the complexity of cancer biology. The ability of OmicsFusionNet to correctly classify advanced cancer stages with the selected features from XGBoost, and deep learning further supports the effectiveness of these features as strong biomarkers. These biomarkers are invaluable for early detection and have significant implications for personalized treatment strategies. Compared to DeepCues developed by Zeng,[22] the highest accuracy among these methods was achieved by our approach. Notably, a remarkable 99.6 % accuracy in pancancer classification was achieved by OmicFusionNet using supervised learning. It can be concluded that RNA-seq data alone are considered effective for pancancer classification, particularly when balanced datasets are utilized with OmicsFusionNet. However, the highest accuracy was achieved by the combination of RNA-seq with methylation data, highlighting the effectiveness of OmicsFusionNet in the integration of different data types and the combination of machine learning with deep learning techniques.

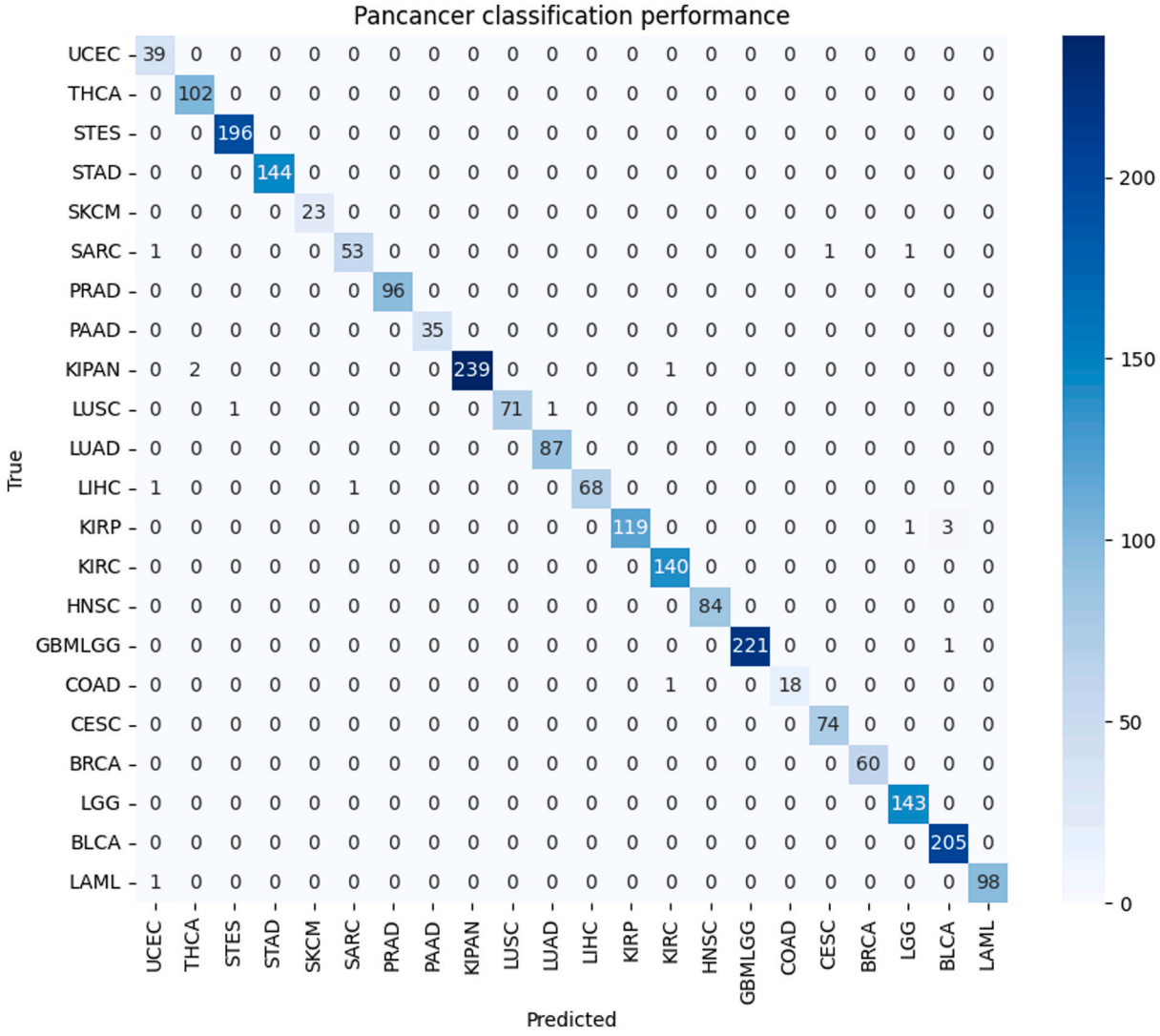


Fig. 7. Confusion Matrix of RNAseq and Methylation data 22 Pancancer classification.

3.3.2. Enrichment analysis and KEGG pathway

The selected features utilized by the OmicsFusionNet model to achieve high classification accuracy comprise 5,857 RNA selected features; Sup. Table S1, and 402 methylation selected features; Sup. Table S2, which were the output of the selecting features phase. These genes have the potential to serve as biomarkers for 22 cancer types. Additionally, enrichment analysis was performed for these features using Enrichr, a comprehensive gene set enrichment analysis web server.[41] Figs. 9 and 10 present KEGG pathways chart graph for RNAseq and methylation-selected genes, respectively.

3.4. Ovarian cancer stage classification results

For ovarian cancer stage classification, a comprehensive analysis was conducted during the feature selection and classification phases to achieve optimal integration of ML and DL, aiming for high accuracy in predicting ovarian cancer stages three and four. This study included five distinct experiments using the CPTAC-OV and OV TCGA datasets, focusing on the integration of multiomics data and employing different models from OmicsFusionNet. In each experiment, identical preprocessing steps were applied, including data cleaning, filtration, and imputation of missing values, and the results were evaluated using accuracy metrics. For all the experiments, the training and testing data were split into 80 % training and 20 % testing sets. The number of samples used in each experiment corresponds to the patient IDs that have data available for all the integrated omics.

3.4.1. Experimental setup

The experiments were conducted using different combinations of omics datasets. Table 4 provides an overview of the omics datasets integrated in each experiment. In each experiment, 12 trials were conducted to explore various combinations of ML and DL techniques for feature selection and classification. ElasticNet was used for feature extraction with fixed parameters (alpha = 0.0005, l1_ratio = 0.2). Each trial combined ML and DL models to determine the optimal number of features and the most effective classification approach.

Five experiments were conducted based on the integration of different omics datasets to evaluate the effectiveness of multiomics integration. In each experiment, 12 trials were performed using OmicsFusionNet hybrid models. The trials were designed to measure the performance of each model and assess the strength of the omics integration. As a result, a total of 60 accuracy measurements were obtained, providing a comprehensive comparison of the models and integration strategies across various omics datasets. Table 5 presents the trial setup and each experiment performance using accuracy metric.

3.4.2. Performance analysis for OmicsFusinNet proposed models

Analyzing the trials' accuracies provides a direct indicator of

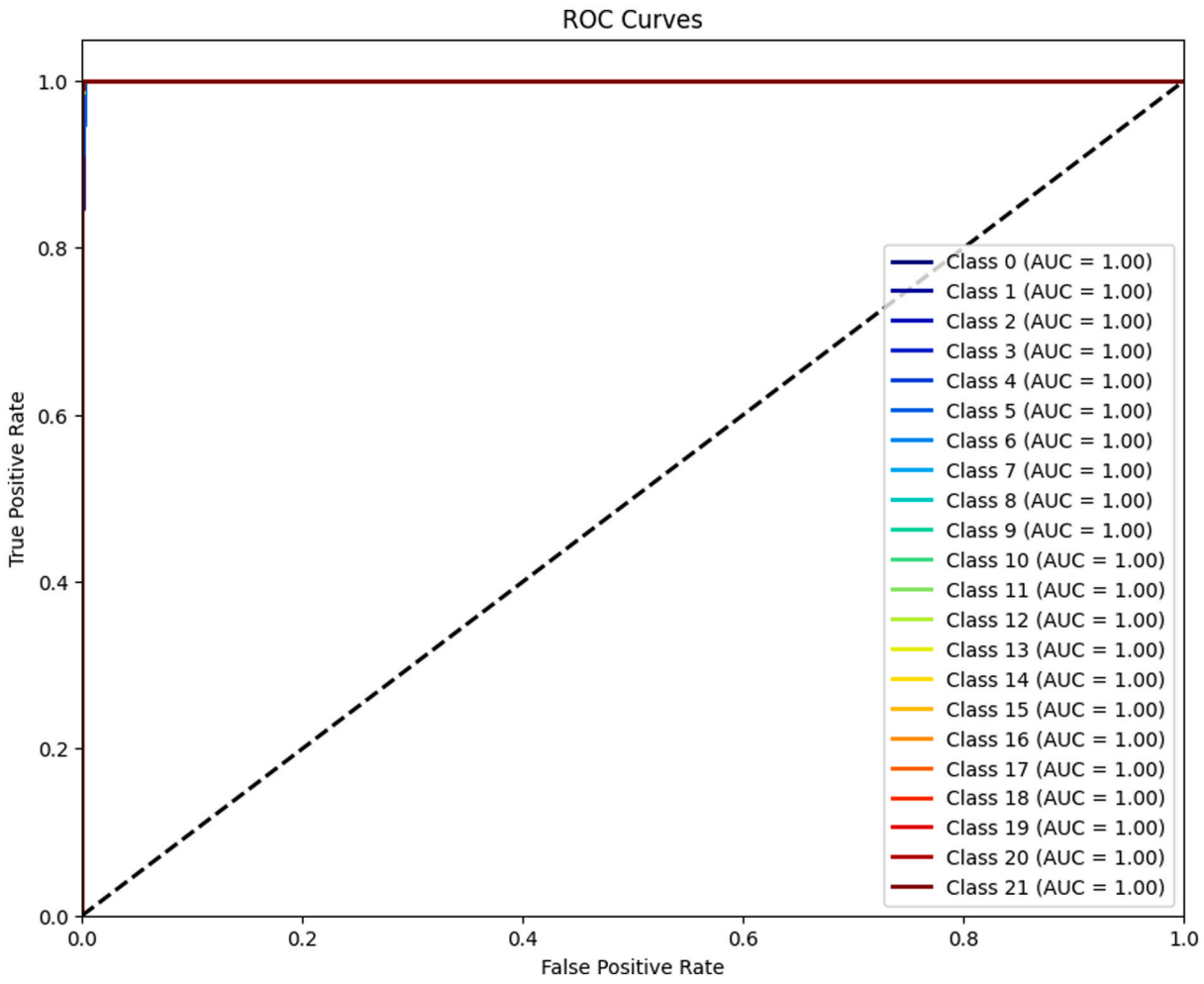


Fig. 8. ROC Curves for Pan-Cancer Classification Across 22 Cancer Types.

OmicsFusionNet performance because it reflects how well the model adapts to varying data inputs, feature selection strategies, and classification techniques, offering a comprehensive view of the model's effectiveness in predicting outcomes. To visualize the performance of OmicsFusionNet models used in ovarian cancer stage classification, a scatter plot with a trend line was plotted in Fig. 11, showing the progression of the test accuracy across the 12 trials.

The trend line in the scatter plot represents the general trend in model accuracy across all trials. It reflects how accuracy changes as we move from Trial 1 to Trial 12. The trend line indicates a slight upward trajectory, meaning that overall performance improves as trials progress. While Trial 7 shows an outlier with lower accuracy, the trend line suggests that this deviation can be accepted as an anomaly. The most significant improvements in accuracy begin from Trial 8 to Trial 12, where the models show consistently higher performance. Overall, the trend line confirms a positive progression in performance across the trials, indicating that model refinements and feature selections are yielding better results over time.

The results in Fig. 11 indicate that the highest accuracy in cancer stage classification was achieved by trials 8 to 12, which were based on OmicsFusionNet Models 1,3, 4, 5, and 6. These models, which combined various feature selection and classification methods, including ElasticNet, XGBoost, SVM, and DL, were consistently shown to perform well.

3.4.3. Deep analysis for OmicsFusinNet proposed models

For in-depth analysis, the three best models were selected: Model One, Model Three, and Model Six. These models achieved high rankings on the trend line. The ROC curves across experiments reveal the varying performance of Model one, Model three, and Model Six. In each experiment, the AUC is used to quantitatively assess how well each model distinguishes between classes. Fig. 12 shows that in Experiment One when used model One achieved the best classification accuracy, but with Models Three and Six performing slightly lower.

In Experiment Two, all three models demonstrated strong performance. Model Six performing the best. So, the integration of ElasticNet and XGBoost enhanced feature selection and accuracy for Model Six. In Experiment Three, classification accuracy improved across all models, likely due to the addition of omics data, with Model Six again performing slightly better. Experiment Four saw further improvements as more omics data were added, with Model One showing slightly higher results, as indicated by its steeper ROC curve. Finally, in Experiment Five, Models One and Six both performed well, with Model Six showing the most balanced and robust accuracy across datasets and feature dimensions. Overall, the ROC curves suggest that Model Six (ElasticNet + XGBoost + DL) consistently performed well across all experiments, especially with larger, complex datasets, while Model One (XGBoost + DL) also showed reliable performance, particularly with multiple omics datasets.

Following the presentation of the ROC curves for Models One, Three, and Six across the five experiments, a confusion matrix was introduced these models but for the first experiment only OV-CPTAC. Fig. 13 validate the results and examine how the test cases are classified by the top three models. This analysis provides a clearer view of the model performance, highlighting specific cases of true positive, false positive, and

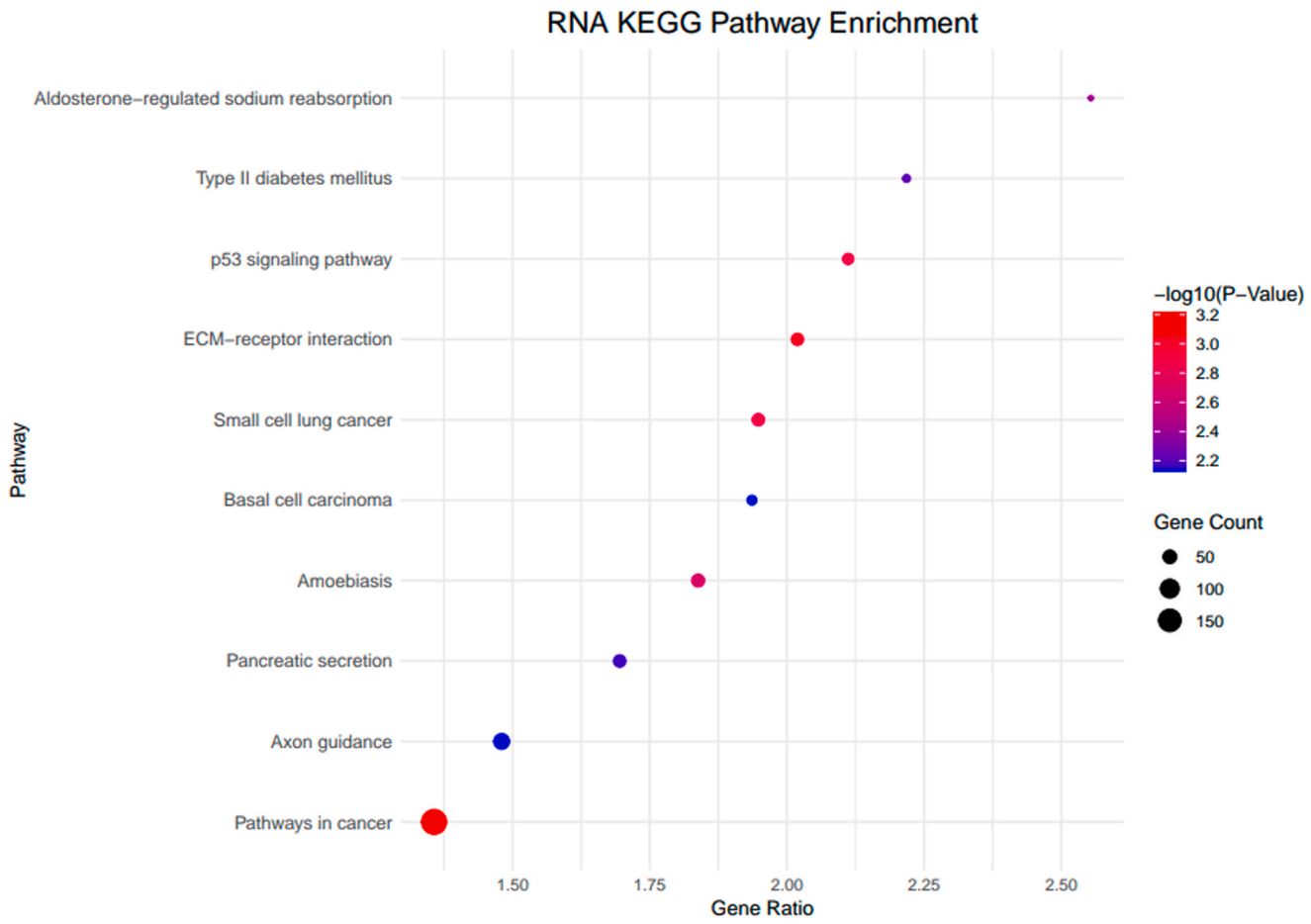


Fig. 9. KEGG Pathway Bar Chart: Pancancer selected RNA Genes.

false negative predictions, which allows for an in-depth understanding of classification accuracy across experiments. By visualizing these confusion matrices, we aim to reinforce the consistency and reliability of our models in differentiating between the cancer types in the test cases.

The confusion matrix was generated on the test cases of the OV-CPTAC dataset, which includes 23 cases: 18 in stage three and five in stage four. From the confusion matrix, it is evident that Model Six is the most efficient, accurately predicting all stage three cases with only two misclassifications in stage four.

The integration of ElasticNet, XGBoost, and deep learning in our models demonstrates a strong performance in Ovarian cancer stage classification. Moreover, the selected features obtained through ElasticNet, and XGBoost play a critical role in identifying strong biomarkers for stage three and four classifications. These biomarkers not only enhance the model's ability to distinguish between advanced stages but also hold significant implications for treatment planning. The reliable identification of these biomarkers could facilitate earlier intervention and potentially inform more effective treatment strategies, underscoring their importance in precision oncology.

3.4.4. Omics integration analysis for ovarian cancer stage classification

To measure the performance of combining different omics data, the ROC curves indicate that Model Six (ElasticNet + XGBoost + DL) which was trial number twelve in the experiment's trails Table 5, consistently demonstrates strong performance across all experiments, especially when handling larger and more complex datasets. Due to its effectiveness in integrating omics data and maintaining robust classification accuracy, Model Six has been selected to use in omics integration analysis.

In the first experiment, RNA, proteome, and phosphoproteome data were used. Strong performance was observed, likely due to the balanced integration of transcriptional (RNA) and post-translational (proteome and phosphoproteome) data, which contributed to effective classification. In the second experiment, the same omics data types were used, but from the TCGA dataset. Performance was slightly lower compared to the first experiment, likely due to differences in data quality and preprocessing between the CPTAC and TCGA sources, highlighting the impact of the data source.In the third experiment, RNA, mutation, SCAN, and methylation data were included. Moderate performance was observed, suggesting that the addition of mutation and methylation and remove proteome data did enhance classification. The model seemed gave the same accuracy as experiment two. In the fourth experiment, RNA, mutation, SCAN, methylation, and proteome data were included. Performance was poor, likely due to the introduction of noise and complexity from multiple omics types, which overwhelmed the model and hindered its ability to classify effectively.

In the fifth experiment, the most comprehensive set of omics data was used, including RNA, mutation, SCAN, methylation, proteome, and phosphoproteome. Despite the inclusion of more data, performance remained moderate, indicating that the additional omics types introduced redundancy and noise, which did not improve the model's effectiveness. As a result, the integration of RNAseq, proteomics, and phosphoproteome data is considered powerful for determining cancer stages, while integrating RNAseq, methylation, mutation, and SCAN data can yield equivalent information.

In contrast to Jeong[29] and Munetoshi,[30] our model demonstrated classification test accuracies ranging from 83 % to 91 %, reflecting its robust performance across comprehensive studies comprising more than

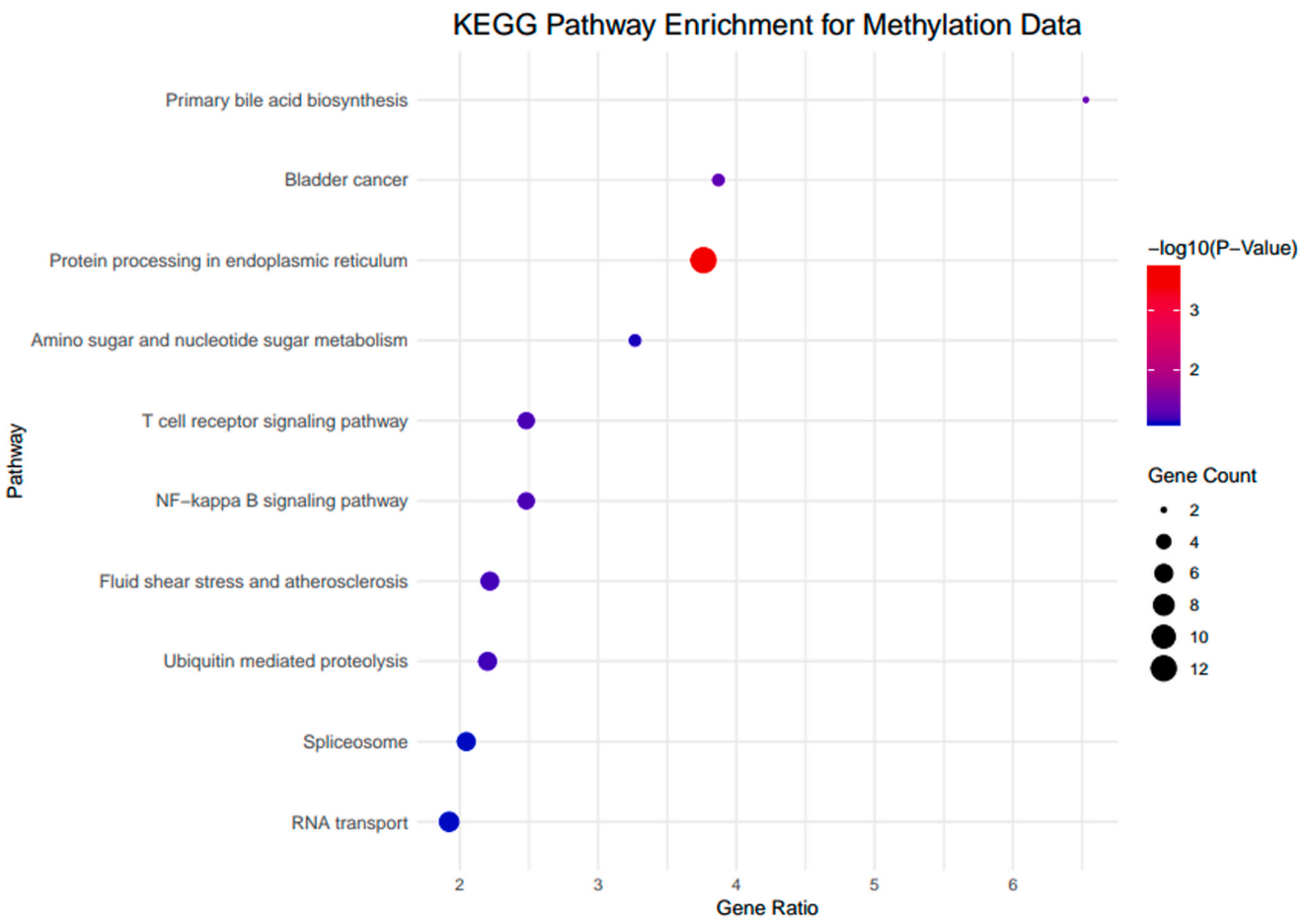


Fig. 10. KEGG Pathway Bar Chart: Pancancer selected Methylation Genes.

Table 4
Summary of Five Experiments on Ovarian Cancer.

Experiment	Omic Datasets	Number of Samples
1st	CPTAC: RNA, Proteome, and Phosphoproteome	75
2nd	TCGA: RNA, Proteome, and Phosphoproteome	17
3rd	TCGA: RNA, Mutation, SCAN, and Methylation	228
4th	TCGA: RNA, Mutation, SCAN, Methylation, and Proteome	63
5th	TCGA: RNA, Mutation, SCAN, Methylation, Proteome, and Phosphoproteome	16

60 trials over five experiments. This extensive evaluation underscores the strength and reliability of our approach for cancer stage prediction through integration strategies. This highlights the efficacy of the proposed approach in predicting ovarian cancer prognosis, demonstrating superior performance over traditional models.

3.4.5. Enrichment analysis and KEGG Pathway

The OmicsFusionNet Model 6, which combines ElasticNet, XGBoost, and deep learning, was found to achieve the highest stage classification accuracy for experiments one, two, and three. In Experiment Three, four omics types—RNA expression, DNA methylation, somatic mutation, and

Table 5
Overview of the 12 Trials in Each Experiment for Ovarian Cancer.

Trial	Models	Classifier	NO. Selected Features	Experiments				
				1st	2nd	3rd	4th	5th
1st	Two	SVM	Top and down 10 ElasticNet	0.83	0.86	0.84	0.89	0.83
2nd	Two	SVM	Top and down 20 ElasticNet	0.91	0.86	0.86	0.84	0.83
3rd	Two	SVM	Top and down 30 ElasticNet	0.83	0.86	0.83	0.84	0.83
4th	Two	SVM	ElasticNet	0.87	0.86	0.84	0.84	0.83
5th	Three	DL	Top and down 10 ElasticNet	0.91	0.86	0.83	0.84	0.83
6th	Three	DL	Top and down 20 ElasticNet	0.87	0.86	0.84	0.84	0.83
7th	Three	DL	Top and down 30 ElasticNet	0.83	0.83	0.83	0.74	0.84
8th	Three	DL	ElasticNet	0.91	0.86	0.86	0.84	0.83
9th	Four	SVM	XGBoost	0.91	0.86	0.86	0.84	0.83
10th	One	DL	XGBoost	0.91	0.86	0.86	0.84	0.83
11th	Five	SVM	Top and down 10 ElasticNet + XGBoost	0.91	0.86	0.86	0.84	0.83
12th	Six	DL	Top and down 10 ElasticNet + XGBoost	0.91	0.86	0.86	0.84	0.83

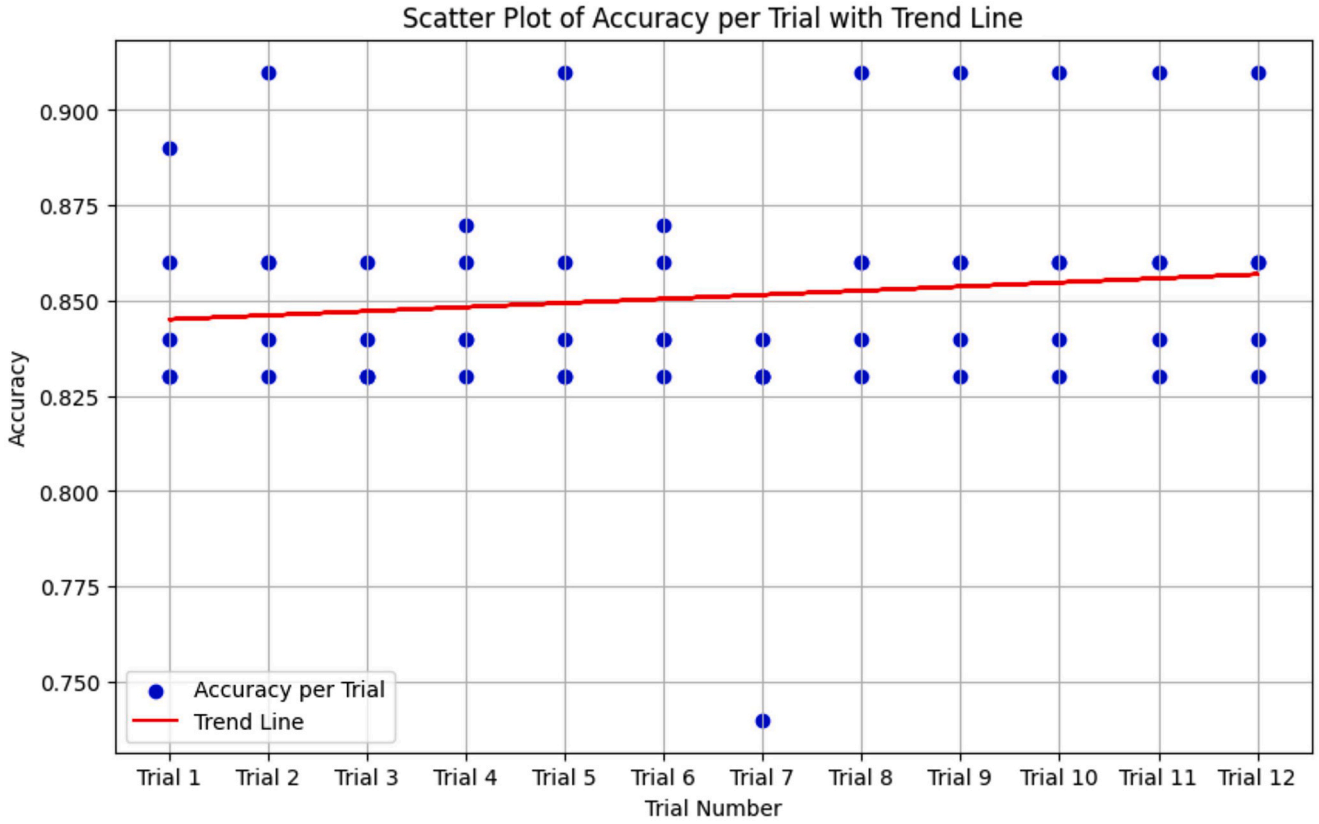


Fig. 11. Scatter Plot of Accuracy Across 12 OmicsFusionNet Trials on Ovarian Cancer Datasets With Trend Line.

SCAN—were utilized. Feature selection was performed using two models. The ElasticNet model was applied to select 20 features per omic, while the XGBoost model identified 98 RNA features (Sup. Table S3), 249 methylation features (Sup. Table S4), 58 mutation features (Sup. Table S5), and 283 SCAN features (Sup. Table S6). These selected features may serve as potential biomarkers for distinguishing between different stages of ovarian cancer. For enrichment analysis, the Enrichr website was employed. Figs. 14, 15, 16, and 17 present the KEGG pathway analysis for RNA, methylation, mutation, and SCAN features selected by ElasticNet and XGBoost, respectively.

4. Conclusion

In this study, comprehensive experiments were conducted to evaluate the integration of ML and DL for pancancer classification and ovarian cancer stage prediction, using RNAseq and methylation data from 23 cancer types. Three experiments were performed with the OmicsFusionNet model. Initially, lower accuracies were observed with unbalanced datasets 80.2 %. However, when the data was balanced; 200 sample per cancer type, 90.6 % classification test accuracy were achieved, emphasizing the importance of balanced datasets. In the final experiment, where RNAseq and methylation data were combined, an impressive accuracy of 99.8 % was reached, highlighting the effectiveness of XGBoost for feature selection and the superior performance of the DL model.

In the ovarian cancer case study, five experiments were conducted using the OV CPTAC and OV TCGA datasets. Six models were applied by OmicsFusionNet, with the combination of ElasticNet and XGBoost features resulting in the highest classification accuracy for cancer staging. Across 60 trials, adaptability to various omics data types, feature selection strategies, and classification techniques was demonstrated by OmicsFusionNet. An upward trend in accuracy was observed from Trial 8 onwards, with Model Six (ElasticNet + XGBoost + DL) consistently yielding the best performance. This model was selected for its strong classification accuracy and effective integration of multiple omics data, particularly in large and complex datasets.

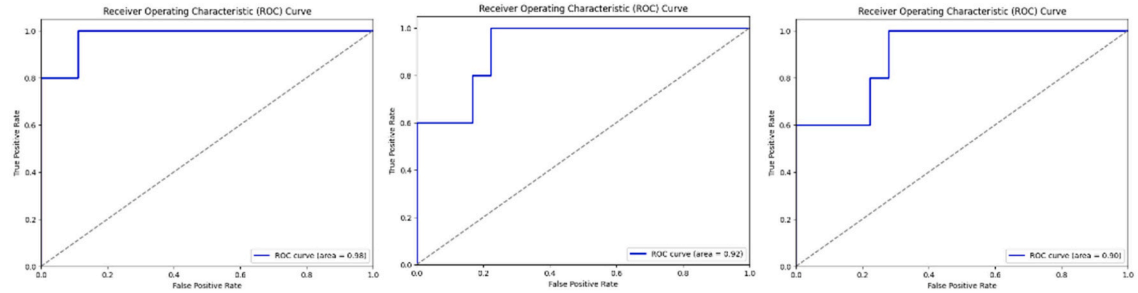
The inclusion of additional omics data generally improved performance, though excessive data occasionally introduced noise and complexity. The importance of integrating RNAseq, proteome, and phosphoproteome data for achieving robust classification results was emphasized in the study. These combined data types provided a balanced view of transcriptional and post-translational processes, significantly improving the accuracy of cancer stage classification. KEGG pathway analysis was applied to the selected features through the integration of XGBoost and ElasticNet, identifying key cancer-related pathways and advancing personalized treatment.

In future work, the number of pancancer types will be increased, and the staging models will be applied to different cancer types. Additionally, pathway analysis may be conducted for the selected genes and proteins to explore the underlying biological mechanisms and their roles in cancer progression.

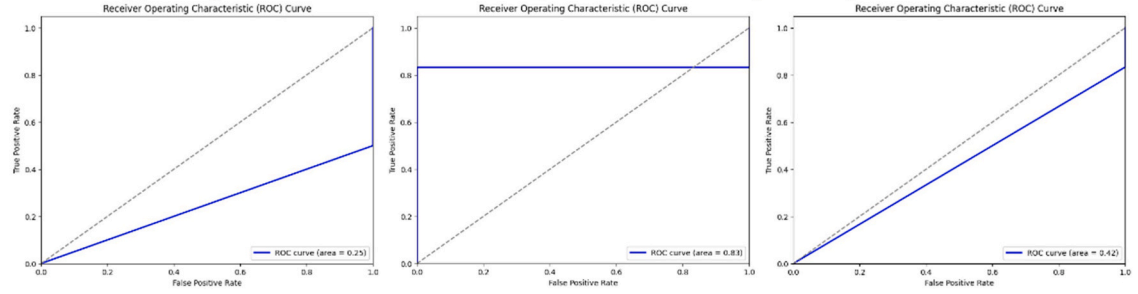
5. Limitations and future work

While a novel multi-omics framework, **OmicsFusionNet**, was introduced for pan-cancer classification and ovarian cancer stage detection, several limitations should be acknowledged, and directions for future work can be identified:

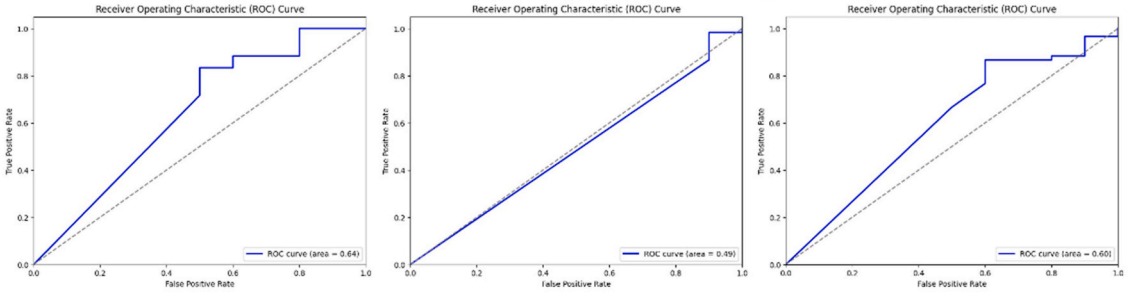
Dataset Diversity and Generalizability:
Although strong performance was observed using publicly available datasets (TCGA and CPTAC), these datasets may not fully represent the diversity of real-world clinical populations or the complete range of cancer types beyond the 23 that were included. Therefore, the generalizability of the model has not yet been assessed on additional cancer types or validated on independent, multi-ethnic cohorts.



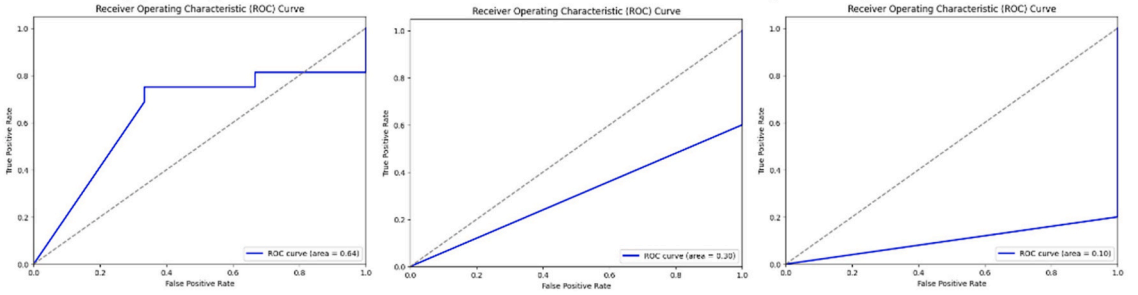
(a) Experiment one for model one, three, and six respectively.



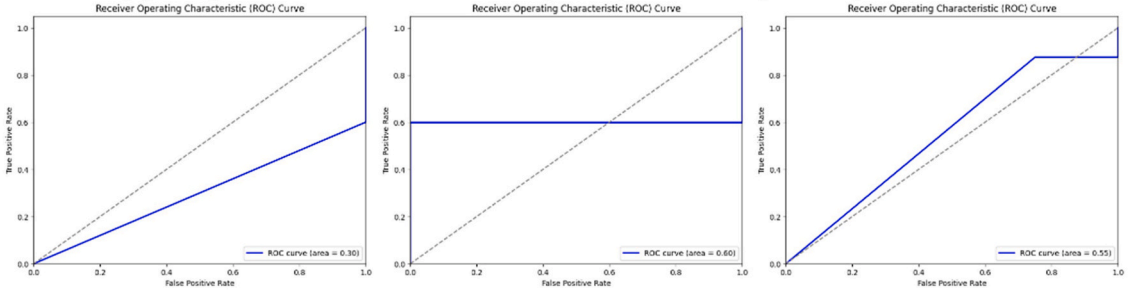
(b) Experiment Two for model one, three, and six respectively.



(c) Experiment three for model one, three, and six respectively.



(d) Experiment four for model one, three, and six respectively.



(e) Experiment five for model one, three, and six respectively.

Fig. 12. ROC Curve Performance Analysis for OmicsFusionNet Models 1, 3, and 6 Across Experiments on Ovarian Cancer.

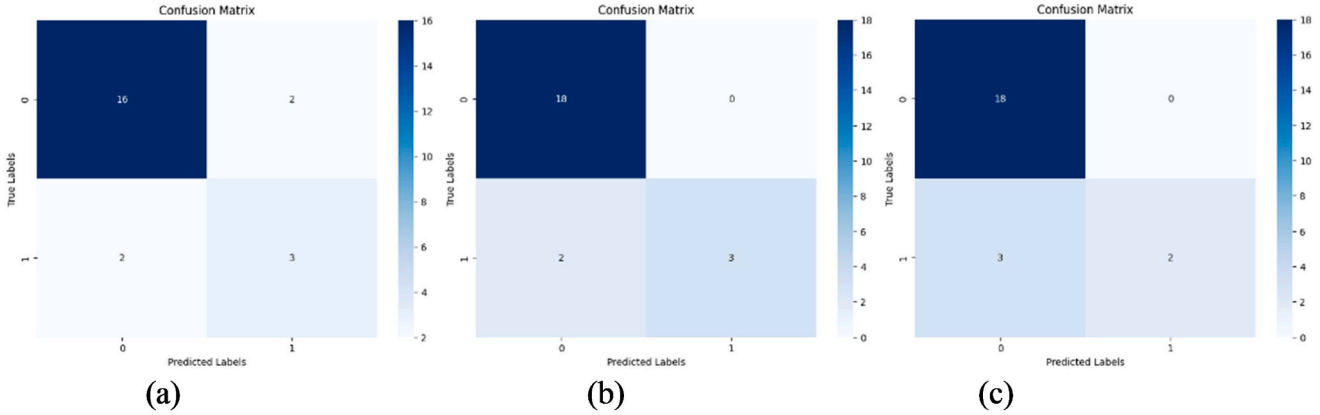


Fig. 13. Confusion Matrices for OmicsFusinNet model one, three, and six over OV-CPTAC dataset. (a) Model three (ElasticNet + Deep Learning)., (b) Model one (XGBoost + Deep Learning), and (c) Model six (XGBoost + ElasticNet + Deep Learning).

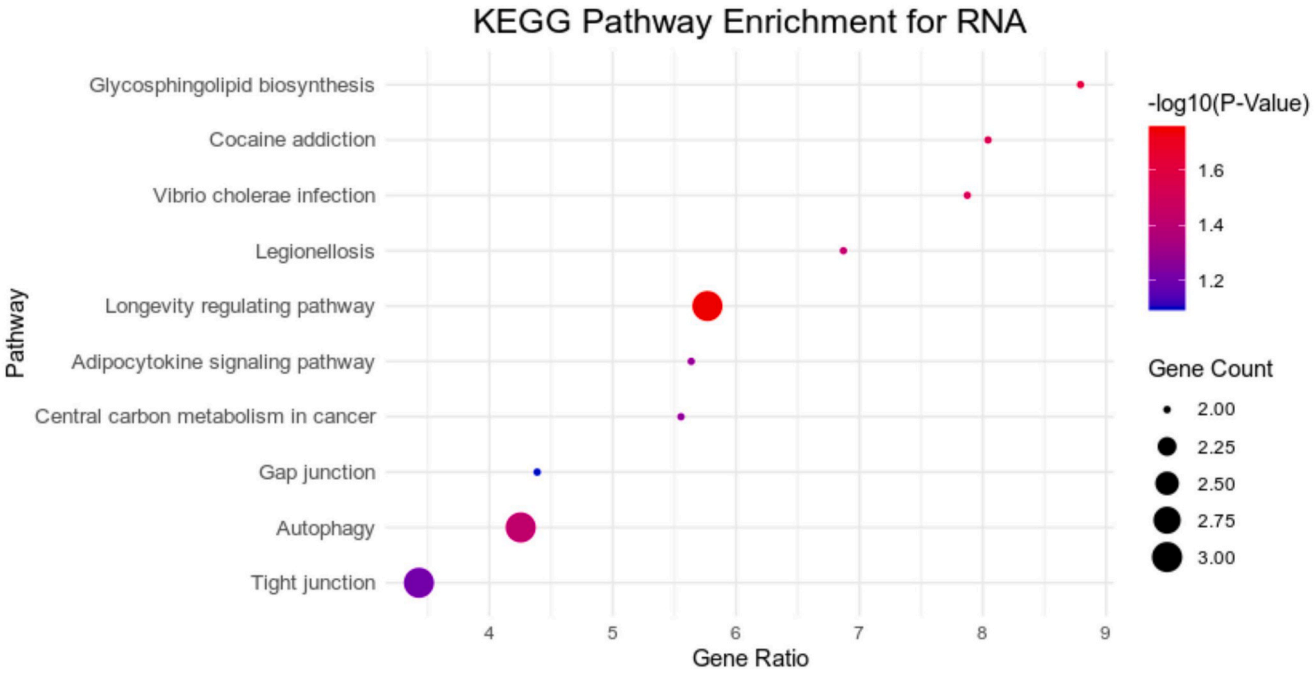


Fig. 14. KEGG Pathway Bar Chart: Ovarian Cancer selected RNA Genes.

Extension of Stage Detection:Stage prediction was applied only to ovarian cancer in this study, serving as a representative case. To validate broader applicability, the approach should be extended to additional cancer types with available stage annotations in future work.

Comparison with Additional Machine Learning Techniques:While ElasticNet and XGBoost were employed, other widely used models such as Random Forest (RF) were not included in the current comparison. A more comprehensive evaluation involving a broader set of machine learning algorithms should be considered in future studies.

Clinical Validation:The proposed model has not yet been validated in a prospective clinical setting. Future work should include validation on real-world clinical datasets or patient-derived samples to assess its practical utility and translational potential.

6. Consent for publication

Not applicable

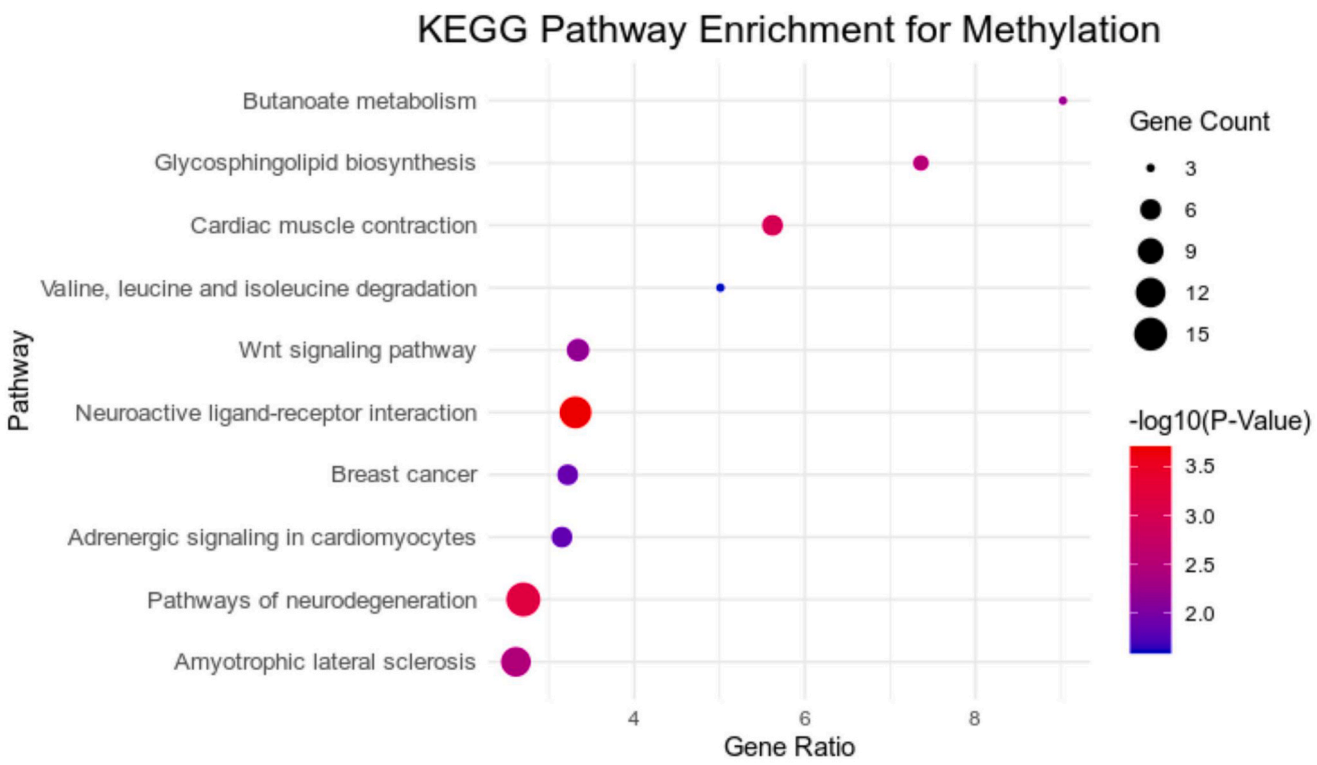


Fig. 15. KEGG Pathway Bar Chart: Ovarian Cancer selected Methylation Genes.

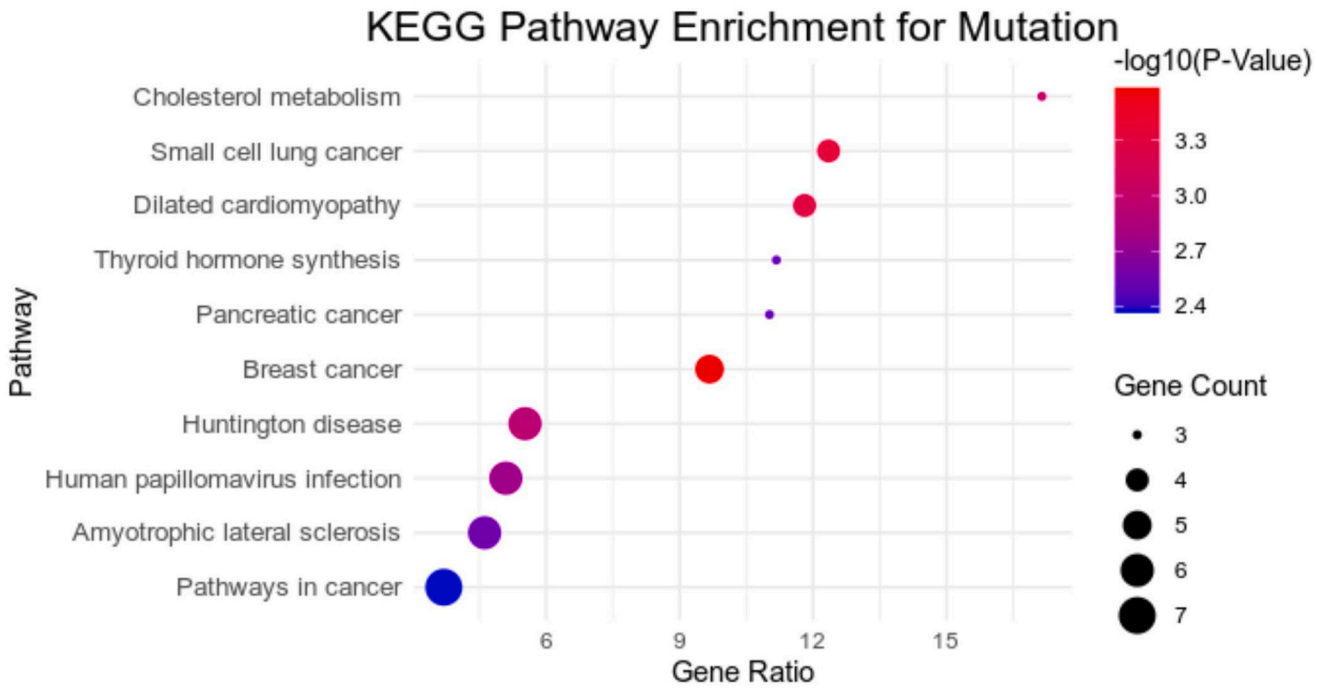


Fig. 16. KEGG Pathway Bar Chart: Ovarian Cancer selected Mutation Genes.

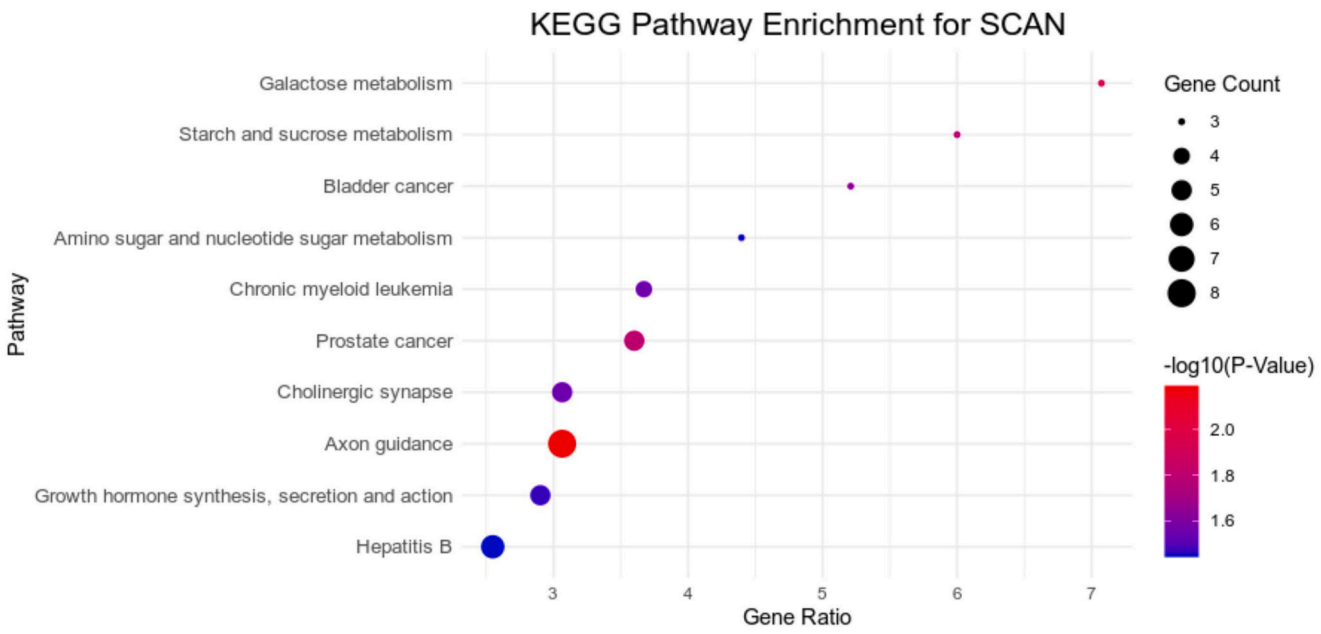


Fig. 17. KEGG Pathway Bar Chart: Ovarian Cancer selected SCAN Genes.

Ethics approval and consent to participate consent for publication

This response clarifies that no new data collection was conducted, so additional ethical approval or consent procedures were unnecessary.

Availability of data and materials

The datasets used and analyzed during the current study are available in publicly accessible repositories. Specifically, data were sourced from The Cancer Genome Atlas (TCGA) https://portal.gdc.cancer.gov/ and the Linked Omics website https://www.linkedomics.org/login.php#dataSource. Detailed information about the datasets utilized in this research can be found in the result section. All data are available from these repositories, and additional details can be requested from the corresponding author upon reasonable request.

CRediT authorship contribution statement

Moshira Ghaleb: Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Maryam Al-Berry:** Writing – review & editing, Supervision. **Hala Ebied:** Writing – review & editing, Supervision. **Mohamed Tolba:** Writing – original draft, Supervision, Project administration.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

I would like to extend my gratitude to my team for their dedication and hard work. I also thank the Faculty of Computer and Information Sciences for their invaluable support throughout this research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jgeb.2025.100639.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics. *A Cancer Journal for Clinicians*. 2020; 70(1):7–30. https://doi.org/10.3322/caac.21590.
2. World Health Organization. Cancer. Retrieved from https://www.who.int/news-room/fact-sheets/detail/cancer. 2022.
3. DeVita VT, Lawrence TS, Rosenberg SA, DePinho RA, Weinberg RA. *"DeVita, Hellman, and Rosenberg's Cancer, "Principles & Practice of Oncology*. 12th ed. Philadelphia, PA: Wolters Kluwer; 2019.
4. Chakraborty S, Hosen MI, Ahmed M, Shekhar HU. "Onco-multiomics Approach, "a New Frontier in Cancer Research. *Biomed Res Int.* Oct. 2018;3. https://doi.org/10.1155/2018/9836256. PMID: 30402498; PMCID: PMC6192166.
5. Murphy KP. *Machine Learning: a Probabilistic Perspective*. MIT Press; 2012.
6. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. MIT Press; 2016. Available at: https://www.deeplearningbook.org/.
7. Ghaleb MS, Al-Berry MN, Ebied HM, Tolba MF. *"Lung Cancer Stages Classification based on Differential Gene Expression and Deep Learning. "10th International Conference on Advanced Intelligent Systems and Informatics (AISI'24)*. 2024.
8. Ghaleb MS, Ebied HM, Shedeed HA, Tolba MF. Image Retrieval based on Deep Learning. *Journal of System and Management Sciences, JSMS*. 2022.
9. Hamed, Esraa Abdelraouf, Tolba, Mohamed, Badr, Nagwa, Salem, Mohammed A.-M., "Large-scale Histopathological Colon Cancer Annotation Model Using Machine Learning Techniques," IJICIS, Vol.23, 3, P:73-82, September 2023.
10. Ghaleb, M.S., Ebied, H.M., Shedeed, H.A., Tolba, M.F., "COVID-19 X-rays Model Detection Using Convolution Neural Network, "International Conference on Artificial Intelligence and Computer Vision (AICV2021). AICV 2021. Advances in Intelligent Systems and Computing, vol 1377. Springer, Cham. 10.1007/978-3-030-76346-6_1.
11. Ghaleb, M.S., Ebied, H.M., Tolba, M.F., "Lung Cancer Stages Classification Based on Differential Gene Expression, "The 3rd International Conference on Artificial Intelligence and Computer Vision (AICV2023), March 5–7, vol 164. Springer, AICV 2023. 10.1007/978-3-031-27762-7_26.
12. Ghaleb, M.S., Ebied, H.M., Tolba, M.F. "Bladder Cancer Microarray Analysis and Biomarker Discovery Using Machine Learning," the 9th International Conference on Advanced Intelligent Systems and Informatics 2023, vol 184. Springer, AISI 2023. 10.1007/978-3-031-43247-7_25.
13. Ghaleb MS, Al-Berry MN, Ebied HM, et al. A novel statistical feature selection framework for biomarker discovery and cancer classification via multiomics integration. *BMC Med Res Methodol*; 2025. https://doi.org/10.1186/s12874-025-02713-z.
14. Ghaleb MS, ElBery M, Moushier H, Tolba MF. ntegrative Radio-Genomic Analysis of Lung Adenocarcinoma: Linking Imaging Features and Gene Expression Profiles for Tumor Classification. *International Journal of Intelligent Computing and Information Sciences*. 2025;25(4):26–39. Available at: https://ijicis.journals.ekb.eg/article_472886.html.

15. Yu XT, Zeng T. Integrative Analysis of Omics big data. *Methods Mol Biol.* 2018;1754: 109–135. https://doi.org/10.1007/978-1-4939-7717-8_7. PMID: 29536440.
16. Jiang X, Li Z, Mehmood A, et al. "A Self-attention Graph Convolutional Network for Precision Multi-tumor Early Diagnostics with DNA Methylation Data. "*interdiscip Sci.* 2023 Sep;15(3):405–418. https://doi.org/10.1007/s12539-023-00563-1. Epub 2023 May 29. PMID: 37247186.
17. Picard M, Scott-Boyer MP, Bodein A, Périn O, Droi A. Integration strategies of multiomics data for machine learning analysis. *Comput Struct Biotechnol J.* 2021 Jun; 22(19).
18. Garcia-Diaz P, Sanchez-Berriel I, Martinez-Rojas JA, Diez-Pascual A, M.",. Unsupervised feature selection algorithm for multiclass cancer classification of gene expression RNA-Seq data. *"genomics*. 2020;112(2):1916–1925.
19. Ramaswamy S, et al. "Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl. Acad. Sci.* 2001;98(26):15149–15154.
20. Lee K, Jeong H-O, Lee S, Jeong W-K. "CPEM: Accurate cancer type classification based on somatic alterations using an ensemble of a random forest and a deep neural network, "Sci. *Rep.* 2019;9(1):1–9.
21. Tabares-Soto R, et al. "A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data, "PeerJ Comput. *Sci.* 2020;6:e270.
22. Zeng Z, Mao C, Vo A. "Deep learning for cancer type classification and driver gene identification. *BMC Bioinformatics*. 2021;22(Suppl 4):491. https://doi.org/10.1186/s12859-021-04400-4.
23. Castillo, D. et al., "Leukemia multiclass assessment and classification from Microarray and RNA-seq technologies integration at gene expression level, " PLoS ONE 14(2), e012127, 2019.
24. Ghaleb MS, Al-Berry MN, Ebied HM, Tolba MF. *"Ovarian Cancer Proteome Analysis and Biomarker Discovery using Machine Learning. "10th International Conference on Advanced Intelligent Systems and Informatics (AISI'24)*. 2024.
25. Xie H, Xu H, Hou Y, Cai Y, Rong Z, Song W, Wang W, Li K., "Integrative prognostic subtype discovery in high-grade serous ovarian cancer, "Journal of Cellular Biochemistry 120(11), 2019, 18659–18666 DOI 10.1002/jcb.29049.
26. Bismeijer T, Canisius S, Wessels LF., "Molecular characterization of breast and lung tumors by integration of multiple data types with functional sparse-factor analysis, "PLOS Computational Biology 14(10), 2018: e1006520 DOI 10.1371/journal. pcbi.1006520.
27. Rappoport N, Shamir R. "NEMO: cancer subtyping by integration of partial multiomic data. "*bioinformatics*. September 2019;35(18):3348–3356. https://doi.org/10.1093/bioinformatics/btz058.
28. Rappoport N, Safra R, Shamir R, "MONET: Multiomic module discovery by omic selection, "PLOS Computational Biology, Vol 16, Issue 9, 2020, 10.1371/journal. pcbi.1008182.
29. Jeong S, Mok L, Kim SI, Ahn T, Song YS, Park T., "Ovarian Cancer Prognostic Prediction Model Using RNA Sequencing Data, "Genomics Inform. 2018 Dec;16(4): e32. doi: 10.5808/GI.2018.16.4. e32. Epub 2018 Dec 28. PMID: 30602093; PMCID: PMC6440672.
30. Akazawa M, Hashimoto K. Prediction of Ovarian Cancer Survival using Machine Learning: a Population based Study. *journal of Women's Health and Development.* 2023;6:68–74.
31. Kaushik AC, Mehmood A, Dai X, Wei DQ. Pan-Cancer Analysis and Drug Formulation for GPR139 and GPR142. *Front Pharmacol.* 2021 Feb;19(11), 521245. https://doi.org/10.3389/fphar.2020.521245. PMID: 33679382; PMCID: PMC7933564.
32. Cortes C, Vapnik, V.. Support-vector networks. *Machine Learning*. 1995;20(3): 273–297.
33. Chen, T., & Guestrin, C., "XGBoost: A Scalable Tree Boosting System, "In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016 New York, NY, USA: ACM. 10.1145/2939672.2939785.
34. Hui Zou, Trevor Hastie, "Journal of the Royal Statistical Society Series B: Statistical Methodology," Volume 67, Issue 2, April 2005, Pages 301–320, 10.1111/j.1467-9868.2005.00503.x.
35. Dong Y, Peng CY., "Principled missing data methods for researchers, "Springerplus,14;2(1):222, 2013 May.
36. Mucherino, A., Papajorgji, P.J., Pardalos, P.M.," k-Nearest Neighbor Classification. In: Data Mining in Agriculture, "Springer Optimization and Its Applications, vol 34. Springer, 2009, New York, NY. 10.1007/978-0-387-88615-2_4.
37. Feng C, et al. "Log-transformation and its implications for data analysis, "Shanghai Arch. *Psychiatry.*. 2014.
38. Hutter C, Zenklusen JC. The Cancer Genome Atlas: Creating Lasting Value beyond its Data. *Cell.* 2018;173(2):283–285. https://doi.org/10.1016/j.cell.2018.03.042.
39. Cancer Genome Atlas Research Network; Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM., "The Cancer Genome Atlas PanCancer analysis project, "Nat Genet. 2013 Oct;45(10):1113-20. doi: 10.1038/ng.2764. PMID: 24071849; PMCID: PMC3919969.
40. Suhas V Vasaikar, Peter Straub, Jing Wang, Bing Zhang, "LinkedOmics: analyzing multiomics data within and across 32 cancer types, "Nucleic Acids Research, Vol 46, Issue D1, P. D956 D963, 4 January 2018, 10.1093/nar/gkx1090.
41. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A. Enrichr: a comprehensive gene set enrichment analysis web server 2016. doi: 10.1093/nar/gkw377. Epub 2016 May 3. PMID: 27141961; PMCID: PMC4987924.