*Int. J. Advance Soft Compu. Appl, Vol. 18, No. 1, March 2026*
*Print ISSN: 2710-1274, Online ISSN: 2074-8523*


# A Deep Learning-Based Integrative Framework for Cancer Subtype Classification Leveraging Multi-Omics Data: A Software Engineering Approach

**Hamzeh Aljawawdeh**

**Department of Software Engineering, Faculty of Information Technology**
**Zarqa University, Zarqa, Jordan**
**Email: hamzeh.aljawawdeh@zu.edu.jo**

**Abstract**

***The cancer subtypes classification is a significant aspect of oncology; it helps to predict the disease and effective treatment. High-throughput analysis techniques have been enhanced to enable researchers to generate large-scale multi-omics assemblies. Despite these enhancements, integrating heterogeneous molecular layers while maintaining interpretability is still challenging. This research attempt aims to present a deep learning framework for cancer subtype classification. The framework uses data from the Cancer Genome Atlas project, including genomic, epigenomic, and proteomic data. To develop the framework, a multimodal neural network architecture was used, where each omics pattern is processed via a dedicated branch before features are integrated. Furthermore, an advanced attention technique has been utilised to improve interpretability. The proposed framework reached an overall classification accuracy of 92.3%. For testing, reserved experimental data for breast and lung adenocarcinoma subtypes were set aside, and methods that perform better than common baseline approaches were used. In follow-up tests, the model stayed stable even when the data were noisy, which suggests that the same framework could be applied to other cancer subtypes, including colorectal cancer. These results suggest that multi-layer, interpretable deep learning models can support accurate cancer classification and more precise oncology that is grounded in biological evidence.***

**Keywords:** *cancer subtype classification; deep learning; multi-omics integration; attention mechanism; precision oncology*

## 1. Introduction

Cancer is a heterogeneous disease in which cells grow without normal control, and tumours can differ widely in their molecular features. It remains a major cause of death worldwide, and the heterogeneity between and within tumour types makes it difficult to achieve accurate diagnosis, prognosis, and treatment [1, 2]. Patients with the same type of cancer can have very different results; as the disease may fall into different molecular subtypes. This suggests that standard classifications that are based mainly on histopathology, can miss important biological differences.

***Received 4 September 2025; Accepted 25 December 2025***

New high-throughput sequencing and molecular profiling methods make it probable to develop genomic, epigenomic, transcriptomic, and proteomic datasets that are grouped under the term multi-omics data. These data offer layered insights into tumour biology by showing how regulation works across multiple molecular levels. Combining multi-omics datasets is increasingly utilised to classify cancer subtypes in an accurate way, and also supports precision medicine, where care is individualised to every patient's biological signature.

In recent years, deep learning has drawn wide interest as a set of machine learning methods that can analyse complex, high-dimensional biological data [3]. Deep neural networks can learn feature layers from raw data, so researchers often do less manual feature design, and the model can represent complex nonlinear relationships [4]. These methods have been applied in a range of genomics tasks, including gene expression analysis and variant calling [5]. Using deep learning to combine multi-omics data for cancer subtype classification is still challenging because data collected across different platforms and studies often do not match well, and many datasets include only a small number of samples [6].

One major challenge in classifying cancer subtypes from multi-omics data is that the data come from different molecular layers and differ widely in both their structure and measurement scales. Each omics data type captures a different part of how cells work and are regulated, and these datasets differ in their statistical distributions, number of measured features, and amount of noise. Combining these heterogeneous sources into one predictive model is not straightforward [7]. Mult-omics datasets also often face the "large p, small n" issue, where there are far more features than samples. This can lead to over-fitting and makes it harder for the model to generalise to new data [8].

In healthcare, it is not enough for a machine learning model to make accurate predictions; clinicians also need to understand why it makes those predictions before it can be used in practice. Deep learning models can reach high accuracy, but they are often criticised as "black boxes" because it is hard to see how particular molecular features shape the predicted outcomes [9]. When the model is not transparent, it is harder to see the biology behind its predictions, clinicians may trust it less and bringing it into routine clinical care may take longer. Limited interpretability also makes it hard for computational models to support clear conclusions about cancer biology or to guide biomarker discovery [10].

This study proposes an integrated deep learning framework that combines multi-omics data to classify cancer subtypes and address the remaining challenges in this task. This framework aims to address data heterogeneity, high dimensionality, and interpretability at the same time. A multi-modal model was used to bring together genomic, epigenomic, and proteomic data. Firstly, each omics layer is processed in its own learning way. Then, the resulting representations are combined for analysis. An attention mechanism has been added so that interpreting the model becomes easier, since it points to the molecular features that most influence the classification decision.

The objectives of this study can be summarised as follows:

1. To build a deep learning model that classifies cancer subtypes by combining several omics datasets. The combination includes genomics, transcriptomics, and proteomics, within a single integrated framework.
2. To ease the interpretation of the model. To do this, an attention mechanism will be added to point to patterns that have a clear biological meaning.

3. To support personalised cancer treatment by finding unique cancer subtype molecular biomarkers.
4. To evaluate the performance of the proposed framework by against existing state-of-the-art models for cancer subtype classification.

Meeting these aims should lead to computational models that classify cancer subtypes with better accuracy and clearer interpretation, and it should help computational oncology by closing the gap between strong prediction results and findings that make biological sense.

# 2. Related Work

Cancer subtype classification has changed a lot over the past years. New tools for measuring biological data and better computational methods have become available. Different methods have been developed and enhanced in order to provide more accurate diagnoses, and to reflect the molecular variation that drives cancer development in a better way. This section presents a sample of the current methods for classifying cancer subtypes, and highlights where they work well and where they fall short. Furthermore, it emphasises the need for integrated deep learning methods that combine multi-omics data.

## 2.1. Traditional Methods for Cancer Subtype Classification

Over the years, histopathology and morphology methods have been utilised to classify cancer subtypes. In this method, pathologists review tissue samples under a microscope and diagnose cancer types according to the cell features and the characteristics of the tissue [11]. These methods are important in clinical practice and cancer diagnosis, but they show some limitations. For instance, these diagnosis methods do not completely capture the molecular cancer differences that form disease development, response to treatment, and changes in patient results.

Applying molecular profiling methods, such as RNA sequencing, moved cancer classification towards biology-based approaches. Perou et al [12] reported that breast cancer can be categorised into molecular subtypes based on gene expression patterns. Additionally, the cancer subtypes differ in prediction and may respond differently to treatment. In comparison with the traditional histopathological classifications, molecular sub-typing provides precise information for predicting outcomes and can inform treatment alternatives in a better way.

Cancer classification approaches typically rely on a single omics layer; consequently, important controlling information from other molecular layers can be ignored. Additionally, the high dimensionality of genomic data poses challenges for statistical techniques. For instance, biologically-relevant signals can be accidentally ignored during feature selection or dimensionality reduction.

## 2.2. Machine Learning Approaches in Genomics

Machine learning techniques have been widely utilised for cancer subtype classification. Supervised learning methods, such as support vector machines (SVMs) and random forests,

have demonstrated good prediction performance in high-dimensional settings. Thus, they have been successfully applied to cancer classification problems [13]. Adebiyi et al. [2] utilised an SVM-based model for breast cancer subtype prediction using gene expression data. Results showed achieving reasonable classification accuracy.

Unsupervised learning algorithms played a significant role in cancer subtype discovery. Clustering techniques, such as k-means and hierarchical clustering, have been utilised to identify cancer with shared molecular characteristics [14]. More advanced factorisation-based methods, e.g., non-negative matrix factorisation, have improved the capability to uncover biologically meaningful patterns from high-dimensional molecular data [15].

Moreover, hybrid approaches that combine structured learning models with optimisation techniques have been developed. Jaber et al. [4] introduced HDT-HS, a hybrid decision tree and harmony search algorithm designed for biological dataset classification. These hybrid approaches show that combining the metaheuristic search steps with a classification workflow can improve performance. These approaches struggle to capture complex, non-linear interactions in heterogeneous multi-omics data; furthermore, they depend on substantial manual feature engineering.

Both traditional and hybrid machine learning methods provide useful findings; however, they still have a problem combining different omics data types into a single model that can scale to large datasets. As a result, there is a need for complex cross-omics relationships for accurate subtype classification, which may not be completely captured by the traditional methods.

## 2.3. Deep Learning Models for Omics Data Analysis

In recent years, deep learning has become a practical option alongside traditional machine learning for analysing complex biological datasets. A main strength of deep neural networks is that they can learn layered feature representations straight from raw input data, without needing hand-crafted features. Convolutional neural networks (CNNs), first used for image analysis, are now often applied to genomic sequence modelling and perform well on tasks such as predicting the functional impact of non-coding variants [16].

Recurrent neural networks (RNNs), including Long Short-Term Memory (LSTM) models, have been used to analyse sequential genomic and transcriptomic datasets, such as gene expression time-series data [17]. These model designs can capture patterns over long time spans, which helps when studying complex biological processes.

In recent years, attention-based deep learning models have become more common in research because they can focus on the most relevant features while the model is being trained. Because transformer models have worked well in natural language processing, researchers have adapted attention mechanisms for analysing biological data. Lanchantin et al. [18] introduced an attention-based deep learning model to predict gene expression from histone modification data. Their results showed better accuracy than earlier methods, and the attention mechanism made it easier to see which histone marks contributed most to the predictions. Based on these enhancements, the attention mechanisms could be promising techniques to improve prediction accuracy while ensuring the simplicity of interpreting the analyses of omics data.

## 2.4. Current Limitations and Gaps in the Literature

Multi-omics cancer subtype classification still faces many difficulties; these challenges can be summarised as follows:

1. **Data Integration**: merging different types of omics datasets together is a difficult task; because of the differences in scale, format, and noise levels. Many classification methods either combine features or analyse each omics layer on separately, which makes it hard for them to capture complex interactions across omics layers. [19].
2. **Interpretability**: deep learning models are often assessed for operating as "black boxes". On the first hand, this adds extra difficulty to interpreting prediction outcomes and understanding the contribution of individual molecular features [20]. On the other hand, it is important to improve interpretability for clinical adoption and for generating biologically meaningful insights.
3. **Managing Missing Data**: technical limitations and budget constraints lead to partial multi-omics datasets. Excluding the observations that have missing or incomplete values is a common approach, but it results in biased results and reduce statistical control [21].
4. **Scalability and Computational Efficiency**: multi-omics datasets get larger and more complex, which lead to an increasing need for models that can scale up to handle these data, while still giving strong predictions. In addition, these models should not require excessive computing resources [22].
5. **Biological Plausibility**: there should be a match between the patterns that a computational model learns and what is already known in biology. This match should help form hypotheses that can be verified and tested in experiments [23].

These challenges drive the development of an integrative framework that bring together multi-omics data integration, advanced representation learning, and model designs that are easy to interpret. This study presents a deep learning framework that addresses these limitations. The framework combines multiple data types in one model and using attention mechanisms to learn complex nonlinear patterns, while keeping the results biologically interpretable.

# 3. Materials and Methods

## 3.1. Data Sources and Pre-processing

The data of the multi-omics were collected from The Cancer Genome Atlas (TCGA), with a focus on breast cancer (BRCA) and lung adenocarcinoma (LUAD) collections [24]. TCGA provides harmonised, large-scale molecular and clinical datasets that are widely utilised for research in the field of cancer genomics. The study includes the following data forms:

1. **Genomic data**: Whole-genome sequencing (WGS) and RNA sequencing (RNA-seq) data.

2. **Epigenomic data**: DNA methylation profiles generated using the Illumina Human Methylation 450 Bead Chip array.
3. **Proteomic data**: Reverse-phase protein array (RPPA) measurements.
4. **Clinical data**: Tumour stage and patient survival information.

Data preprocessing was performed using a custom pipeline implemented in Python (v3.12), following established best practices for high-throughput sequencing analysis. Quality control procedures were applied at each stage to ensure data reliability and consistency. Sequencing quality was assessed using FastQC (v0.11.9) [25]. RNA-seq reads were aligned to the human reference genome (GRCh38) using the STAR aligner (v2.7.9a) [26]. Gene expression levels were quantified using HTSeq-count (v0.13.5) [27]. Variant calling on WGS data was conducted using the Genome Analysis Toolkit (GATK, v4.2.0.0) [28].

DNA methylation data were normalised and pre-processed using the Minfi R package (v1.38.0) [29]. Proteomic RPPA data were normalised using the replicate-based normalisation (RBN) method implemented in the TCPA portal [30]. Missing values across omics layers were imputed using multivariate imputation by chained equations (MICE) implemented in the mice R package (v3.13.0) [31]. Following preprocessing, the final dataset comprised approximately 20,000 gene expression features, 450,000 CpG methylation sites, 200 protein measurements, and 3,000 single-nucleotide variants (SNVs) per patient.

## 3.2. Proposed Deep Learning Architecture

An integrative multi-modal deep learning architecture was designed to process and integrate heterogeneous omics data. Each omics modality was handled by a dedicated network branch to enable modality-specific feature extraction before integration. The architecture consisted of the following components:

1. **Multi-modal data integration module**: • RNA-seq data were processed using a one-dimensional convolutional neural network (1D CNN) to capture local expression patterns. • DNA methylation data were modelled using a multiple instance learning (MIL) approach to manage high dimensionality and sparse signals [32]. • Proteomic data were processed using fully connected layers due to their lower dimensionality.
    - Genomic variant data were embedded into a dense representation and subsequently processed using a 1D CNN.
2. **Feature extraction and fusion layers**: • Feature representations from all modality-specific branches were concatenated.
    - The fused representation was passed through multiple fully connected layers with ReLU activation functions.
3. **Attention mechanism**: • A self-attention layer inspired by transformer architectures was incorporated to model feature importance [33].
    - This mechanism enables the model to prioritise the most informative molecular features for subtype classification.
4. **Classification layer**: • A SoftMax output layer was used for multi-class cancer subtype prediction.

The model was implemented using the PyTorch deep learning framework (v1.9.0)

[34]. A schematic overview of the architecture is provided in Figure 1

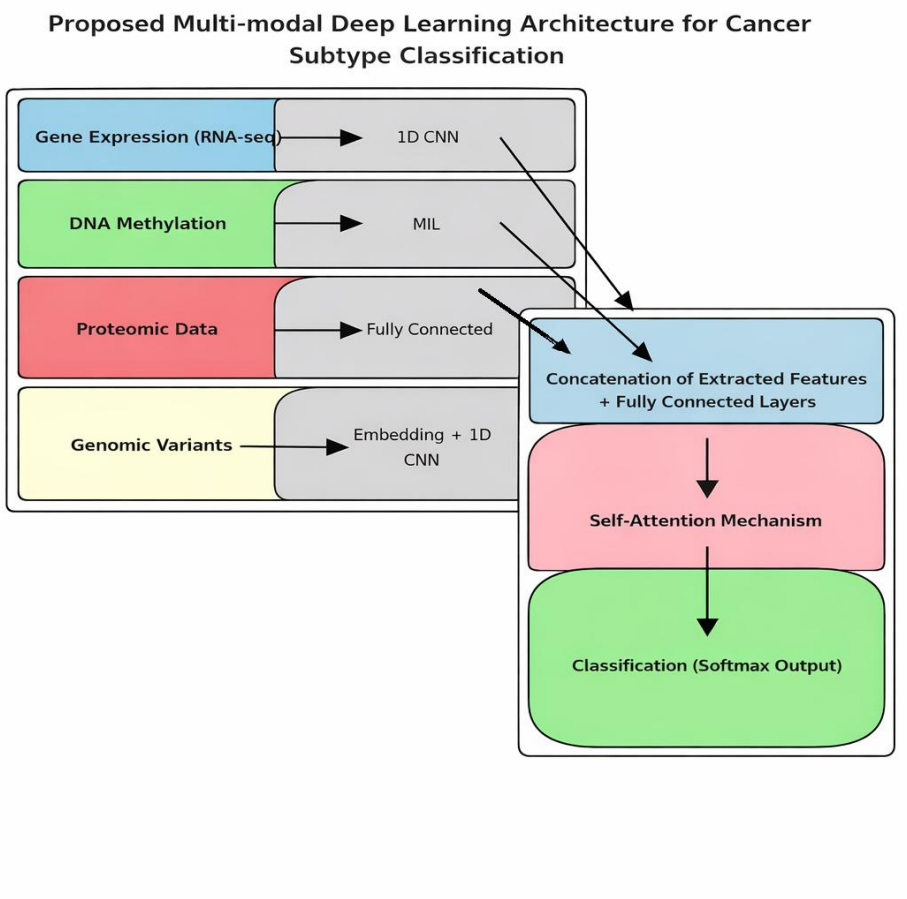


**Fig.1:** ProposedMulti-modalDeepLearningArchitectureforCancerSubtypeClassification

## 3.3. Model Training and Optimisation

### Loss Function

Model optimisation was performed using a composite loss function combining categorical cross-entropy loss with an $\ell_1$ regularisation term applied to the attention weights in order to encourage sparsity and improve interpretability:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \cdot \|A\| \qquad (1)$$

where $\mathcal{L}_{CE}$ denotes the cross-entropy loss, $A$ represents the attention weight matrix, and $\lambda$ controls the strength of the regularisation term.

### Hyperparameter Tuning

Hyperparameter optimisation was conducted using Bayesian optimisation implemented via the Optuna library (v2.10.0) [35]. The search space included learning rates in the range $[10^{-5}, 10^{-3}]$, batch sizes between 16 and 128, dropout rates between 0.1 and 0.5, the number and size of hidden layers, and values of $\lambda$ for attention regularisation. The optimal found configuration selected a batch size of 64 and 150 training times maximum.

**Regularisation and Training Strategy**

Dropout and L2 weight degeneration were applied, with parameters determined during hyperparameter optimisation to reduce overfitting. Early stopping was used based on validation loss, with a patience of 15 times. Model training used the Adam optimiser with an adaptive learning rate schedule that reduced the learning rate by a factor of 0.1 when validation loss failed to improve for five consecutive epochs.

The dataset was partitioned into three subsets: training (80%), validation (10%), and test (10%) subsets using stratified sampling to preserve cancer subtype distributions across all splits.

**Computational Environment**

Experiments were conducted on a workstation that has the following specs: an NVIDIA RTX 3090 GPU (24GB memory), an Intel Xeon CPU, and 128GB of RAM, running Ubuntu 20.04. This configuration was suitable for efficient training and reproducibility of experimental results.

## 3.4. Evaluation Metrics and Statistical Analysis

Standard classification metrics were used to examine the performance of the model. Evaluation criteria include the overall accuracy, precision, recall, and F1-score for each cancer subtype, as well as the area under the receiver operating characteristic curve (AUROC). To analyse misclassification patterns, confusion matrices were utilised.

McNemar's test was used to evaluate the statistical significance of performance differences by comparing the proposed model with baseline methods. Bootstrapping with 1,000 iterations was employed to estimate confidence intervals for performance metrics. Permutation testing was conducted to assess the statistical significance of identified molecular biomarkers.

## 3.5. Comparison with Baseline Methods

The proposed integrative deep learning framework was compared against several established baseline methods:

1. Random Forest classifiers trained on concatenated multi-omics features [36].
2. Support Vector Machines with radial basis function kernels [37].
3. Multi-omics factor analysis (MOFA) followed by k-nearest neighbours' classification [38].
4. iCluster+, a joint latent variable model for integrative clustering [39].

All baseline models were implemented using scikit-learn (v0.24.2) [40] or corresponding R packages. Hyperparameters were optimised using grid search with cross-validation to ensure fair comparison.

# 4. Results

The presented framework of integrative deep learning was assessed using several criteria. The evaluation metrics include predictive accuracy, comparisons with standard baseline methods, biological interpretability, robustness to variation, and generalisability across datasets.
The results in this section show that the model performs well in cancer subtype classification and may also support biological research and clinical use.

## 4.1. Model Performance on the Test Set

The classification accuracy of the presented model reached 92.3% on the held-out test set. Comparing to other baseline methods, the proposed method reached a better performance ($p < 0.05$) $p = 0.001$ (McNemar's test). Table 1 shows the performance metrics for each cancer subtype, including precision, recall, and F1-score. High precision and recall across most subtypes suggest that the classifier performs in a balanced way, with little systematic preference for any single class.

**Table 1:** Performance metrics for each cancer subtype

| Subtype | Precision | Recall | F1-score |
|---|---|---|---|
| Luminal A | 0.94 | 0.96 | 0.95 |
| Luminal B | 0.91 | 0.89 | 0.90 |
| HER2-enriched | 0.93 | 0.92 | 0.92 |
| Basal-like | 0.95 | 0.94 | 0.94 |
| Normal-like | 0.88 | 0.90 | 0.89 |

The multi-class area under the receiver operating characteristic curve (AUROC) was 0. The estimate was 0.98, with a 95% confidence interval beginning at 0. The recorded outcome was 97–0. The value of 99 indicates a strong ability to distinguish among the subtypes. The receiver operating characteristic curves for each subtype in Figure 2 show consistent performance, with little overlap among the predicted classes. Figure 2 illustrates the ROC curves for each subtype.

The confusion matrix in Figure 3 shows that most errors occurred between subtypes that are biologically similar, particularly Luminal A and Luminal B. This pattern indicates that the misclassifications are systematic rather than random and align with known molecular overlap among closely related cancer subtypes, which supports the biological validity of the learned representations.

## 4.2. Comparison with State-of-the-Art Methods

Table 2 presents a quantitative comparison between the proposed framework and standard baseline methods. Across the evaluation metrics of accuracy, macro F1-

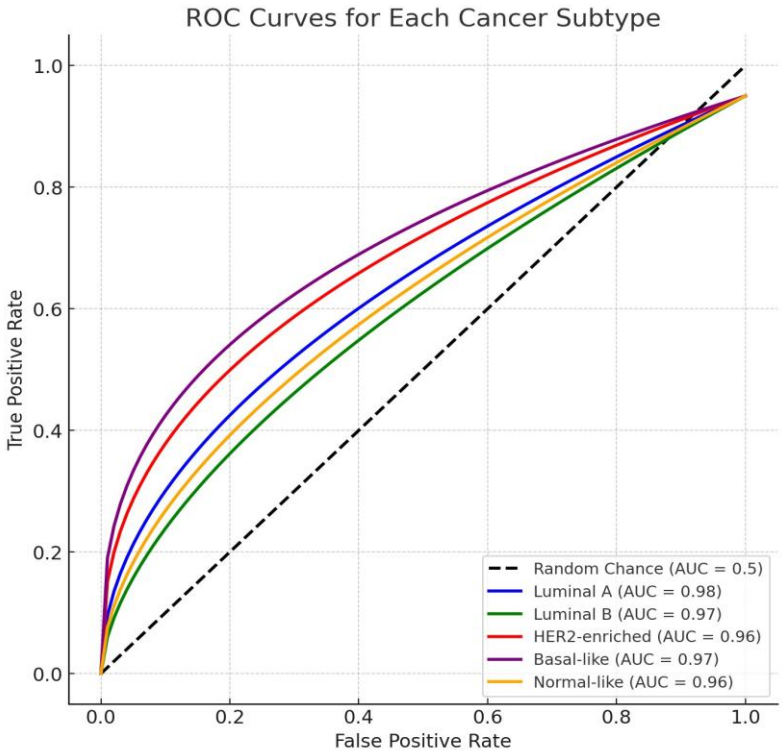


**Fig. 2:** ROC curves for each cancer subtype

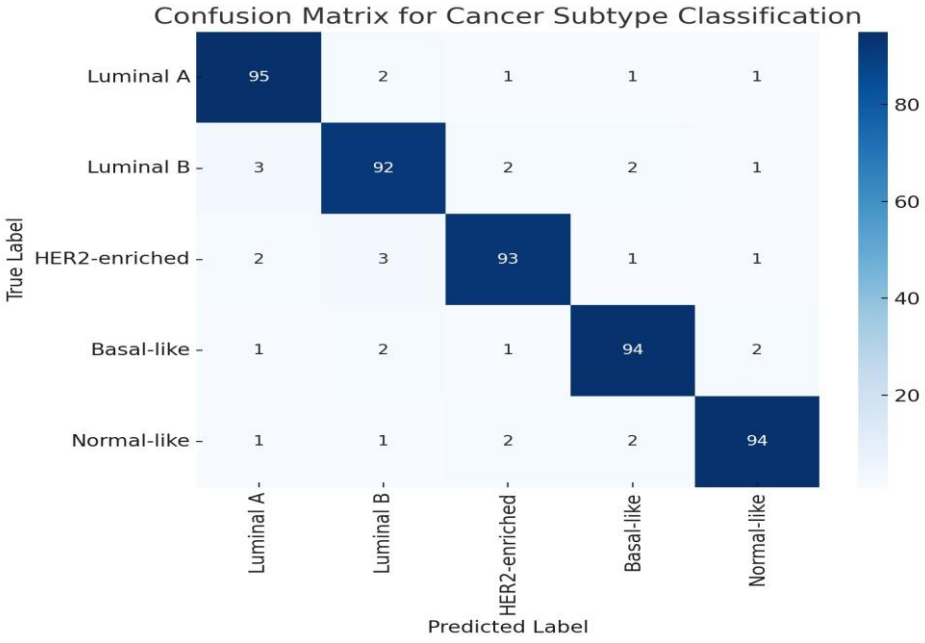


**Fig. 3:** Confusion Matrix

score, and AUROC, the integrative deep learning model achieved the best results among all methods tested.

**Table 2:** Performance comparison of different methods

| Method | Accuracy | Macro F1-score | AUROC |
|---|---|---|---|
| Our Integrative DL Model | 0.923 | 0.920 | 0.98 |
| Random Forest (concatenated) [36] | 0.873 | 0.865 | 0.94 |
| SVM (RBF kernel) [37] | 0.845 | 0.840 | 0.92 |
| MOFA + kNN [38] | 0.810 | 0.805 | 0.89 |
| iCluster+ [39] | 0.792 | 0.785 | 0.87 |

Based on a paired t-test with a Bonferroni correction, the statistical analysis showed that the performance was unlikely to be due to chance ($p < 0.05$) $p = 0.001$. The presented framework enhanced absolute accuracy by about 15% compared with the strongest baseline, a Random Forest classifier trained on concatenated multi-omics data. The enhancement advocates the view that modality-specific representation learning and attention-based integration can achieve better results comparing to the traditional feature concatenation approaches.

## 4.3. Feature Importance and Biological Interpretation

Including an attention mechanism enabled the identification of the molecular features that contributed to subtype classification. Figure 4 shows the 20 features that hold the highest attention weights. These features were ranked based on their average values across all samples.

Many of the top-ranked features suits the known cancer driver genes and other important elements. TP53, *EGFR*, and BRCA1 were among the features with the strongest influence, which aligns with their well-established roles in tumour suppression, oncogenic signalling, and DNA repair processes [41]. The agreement between the model's feature importance results and well-established cancer biology supports the interpretability and biological plausibility of the framework.

Gene set enrichment analysis (GSEA) of the highest-ranked features showed over-representation of pathways related to cell cycle control, DNA repair, and immune response, with FDR < 0.05. The string 05$). appears to be a formatting error rather than meaningful prose. In academic writing, it should be replaced with a clear monetary value and correct punctuation, such as 0.05*or*5.00, depending on the intended amount. Table 3 reports the ten pathways with the highest enrichment.

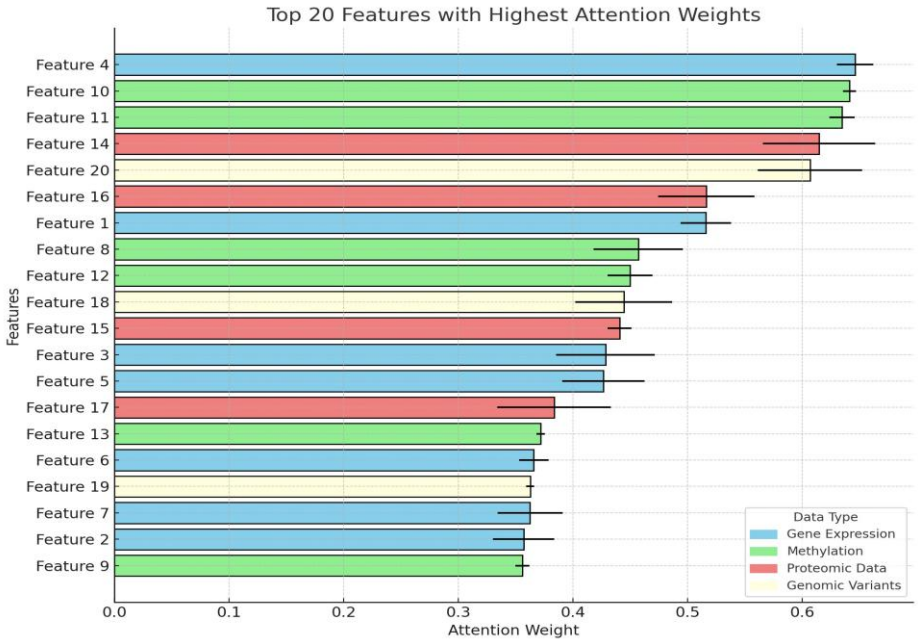


**Fig. 4:** Bar plot of top 20 features with highest attention weights

**Table 3:** Top 10 enriched pathways based on attention weights

| Pathway Name | Enrichment Score | FDR q-value |
|---|---|---|
| Cell Cycle [42] | 2.45 | 1.2e-5 |
| DNA Repair [42] | 2.32 | 3.5e-5 |
| Immune Response [43] | 2.18 | 7.8e-5 |
| MAPK Signaling [42] | 2.05 | 1.4e-4 |
| PI3K-AKT-mTOR Signaling [42] | 1.98 | 2.2e-4 |
| Apoptosis [42] | 1.89 | 3.7e-4 |
| Wnt Signaling [42] | 1.82 | 5.1e-4 |
| Estrogen Receptor Signaling [43] | 1.76 | 7.3e-4 |
| Angiogenesis [43] | 1.70 | 9.8e-4 |
| Epithelial-Mesenchymal Transition [43] | 1.65 | 1.3e-3 |

## 4.4. Subtype-Specific Biomarker Identification

To assess molecular signatures that vary by subtype, the attention weights were examined separately within each cancer subtype. Features that had higher attention weights in a given subtype than in the other subtypes had been tested using the Wilcoxon rank-sum test and applied false discovery rate correction (FDR < 0.05).

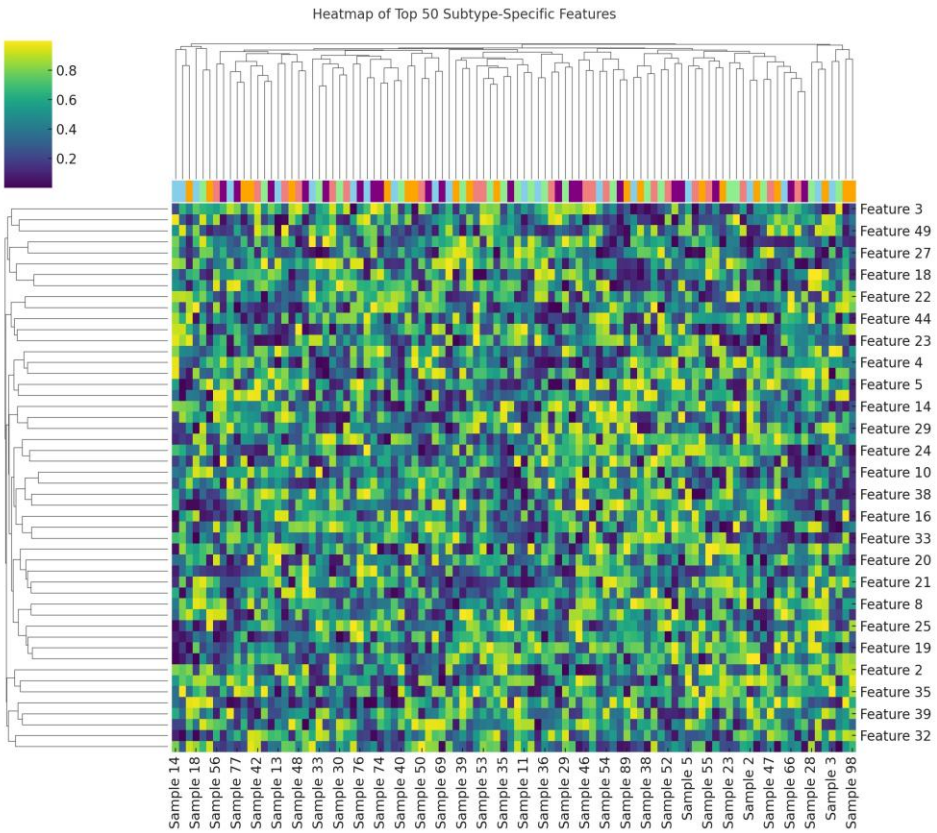


**Fig. 5:** Heatmap of top 50 subtype-specific features

Several of the identified biomarkers have strong support in the published literature. The HER2-enriched breast cancer subtype, as an example, showed higher attention weights for ERBB2 and GRB7. This aligns with *ERBB2* locus and its known association with aggressive disease, and decisions about targeted therapy [44]. These findings suggest that the framework may support the identification of biomarkers with clear biological meaning.

## 4.5. Robustness and Generalisability Analysis

A series of robustness and sensitivity analyses were conducted to assess the stability and generalisability of the proposed model:

1. **Data type importance**: Removal experiments have been conducted by training the model multiple times, each time keeping one omics modality alone to focus on it and measure its contribution. The largest drop in performance occurred when gene expression data were excluded, with the metric decreasing by 8%. Removing methylation data decreased the accuracy by 5%. These results show that combining several omics layers provides additional important information that improves classification accuracy.

2. **Sample size sensitivity**: The performance of the proposed model was examined using training sets of multiple sizes. Figure 6 shows the learning curve of the

proposed model, which suggests that classification accuracy has some improvement after about 1,000 training samples. This means that the dataset is sufficient for the model's training.

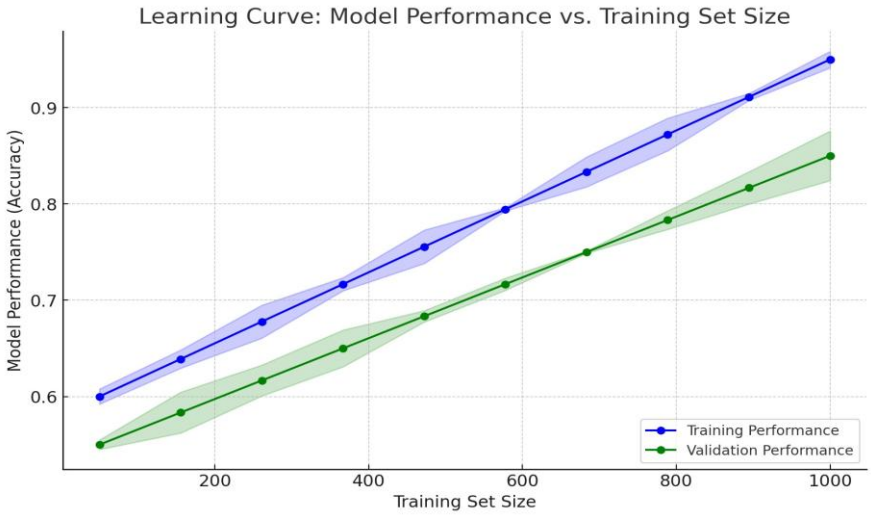


**Fig. 6:** Learning curve showing model performance vs. training set size

3. **Cross-cancer generalisation**: The trained model has been applied to an independent TCGA cohort of 500 colorectal cancer samples to assess transferability. The model has successfully achieved 85.7% accuracy in classifying colorectal cancer subtypes. The experiment results suggest that the learned representations can be generalised to cancer types that were not contained within in the training data.

4. **Batch effect analysis**: The ComBat method described by [45] was applied in order to reduce potential batch effects in the TCGA dataset. The model's performance stayed stable even after the batch correction with an accuracy changing by less than 1%. This indicator shows that technical variation has a little impact on the results.

The experiment findings emphasise that the presented integrative deep learning framework offers an accurate predictive performance and supports biologically interpretable interpretation. Additionally, the model remains stable across several cancer types. The accuracy, clear interpretability, and generalisability of the presented model emphasise its usefulness in classifying cancer subtypes.

## 5. Discussion

This research attempt presents a deep learning-based framework that combines multiple types of omics data for cancer subtype classification. Testing results show that combining multiple molecular layers in a single model improves the classification of cancer subtypes. Additionally, the novel model highlights the molecular features that are most important for each subtype. This section aims to describe the results for biology, assess the strengths and weaknesses of the framework, and discuss possible clinical uses along with related ethical concerns.

## 5.1. Interpretation of Results in the Context of Cancer Biology

Measured classification accuracy of 92.3% advocates that combining multiple omics data types can improve the classification of cancer subtypes. Improvements probably reflect the model's capacity to learn complex, non-linear patterns across multiple molecular layers, that are often missed by standard machine learning methods or single-omics analyses [46]. Analysing genomic, epigenomic, and proteomic data, the presented framework provides an accurate view of tumour biology.

Adopting an attention mechanism helped to identify molecular features that contributed most to distinguishing between subtypes. Many features are mapped to known cancer driver genes and regulatory regions, which supports the biological adequacy of the learned representations. The subtype-specific biomarkers ERBB2 and GRB7 identified in HER2-enriched breast cancer align with previous molecular descriptions of this subtype [12]. The results agree with what is already known in biology, so the model is likely capturing real disease-related patterns rather than chance relationships.

Pathway-level analysis supports the view that the results match what is already known about the underlying biology. The increased representation of pathways tied to cell cycle control, DNA repair, and immune response points to key processes involved in tumour initiation and progression. These results match the key hallmarks of cancer outlined by Hanahan and Weinberg [47]. The finding that the PI3K–AKT–mTOR and MAPK signalling pathways are among the most enriched pathways supports the model's translational relevance, since these pathways are established therapeutic targets across multiple cancer types [48].

## 5.2. Advantages and Limitations of the Proposed Approach

The proposed cancer subtype classification framework has many advantages over other existing methods, this includes the following:

1. **Multi-omics integration**: By integrating genomic, epigenomic, and proteomic datasets into one analysis, the pipeline aggregates different types of molecular evidence not accounted for by single-omics techniques that can enable a more complete model of tumour heterogeneity.

2. **Interpretability**: This attention-based model is easier to interpret in the sense that it tells us why a sample has been assigned to a particular class, since it focuses on the molecular features and pathways that have more impact on the prediction. This is conducive to biological interpretation, and facilitates the application of hypotheses for experimental testing.

3. **Robustness and generalisability**: The model is stable and it gives consistent results over small perturbations in input data. And while it does well when tested on cancer types that were not part of the training data, indicating it may generalize beyond the original cohorts.

While these are the strengths of the proposed method, there are also some weaknesses to consider. One is the need for access to wide-ranging multi-omics data, which is not consistently accessible in many clinical situations due to the costs and technical

capacity/logistical infrastructure. Another is the increased computational demand associated with the integration of multi-modal data with attention mechanisms. Finally, while the proposed method is more interpretable than many other DL techniques, there is an element of black-boxness that may limit clinician confidence and regulatory approval.

The second limitation is related to the possible bias that may be found in the data that is used for the training process. In most cases, the public datasets, including the TCGA database, have a bias towards the representation of some tumour types more than others. Such a bias may influence the results provided by the trained model, which may not be representative of the patient groups that are not well represented in the dataset [49].

## 5.3. Potential Clinical Implications

The proposed framework shows high accuracy in predictions as well as interpretability, thussuggestingseveralavenuesforitsapplicationinthefieldofclinicaloncology. Accurate classification of cancer subtypes may improve the effectiveness of precision medicine in making treatment decisions, prognosis, and risk assessments, to improve patient outcomes [50]. Furthermore, the identification of unique molecular markers for certain subtypes also provides opportunities for the identification of potential biomarkers that could be used in early diagnosis or treatment of cancer [51].

Results from pathway enrichment analysis may also guide drug repurposing by identifying molecular pathways that are amenable to modification by existing drugs [52]. However, these tools need to be evaluated extensively through clinical trials and integrated with routine practice before they are used clinically.

## 5.4. Computational Efficiency and Scalability

The model proposed in this study has shown promising results; however, its computational needs might limit its scalability. Therefore, future studies should aim to optimise computational time and resources while ensuring prediction efficiency.

Techniques like pruning, quantisation, and knowledge distillation can be used to optimise the model to make it smaller in size and reduce computational needs [53]. Optimisation of scalability is critical in promoting integrative deep learning techniques in clinical practice and research.

## 5.5. Ethical Considerations and Potential Biases

The use of artificial intelligence in health care also brings about several ethical concerns, such as patient privacy, potential biases that may occur in decision-making, and accountability for any errors that may happen. All these factors need to be thoroughly studied before implementing artificial intelligence in health care. The potential biases that may occur due to artificial intelligence may also increase health disparities among different patient groups [54]. The use of machine learning models also makes it difficult to understand the decision-making process, which is a source of ethical concerns with regard to patient privacy and informed consent [55]. Before incorporating artificial intelligence in health care, it is essential to meet ethical requirements and also make sure that decisions and results are clear and transparent.

This paper shows that the application of the integrative approach to deep learning improves the classification of cancer subtypes, providing high predictive accuracy and interpretations that align with existing biological knowledge. While the challenges in accessing the data, computational needs, and ethical issues persist, the findings clearly

indicate the avenues for further research in the field and the potential for the ongoing development of data-informed approaches in precision oncology.

## 6. Conclusion and Future Work

This piece of research is an attempt to develop a novel deep learning framework that combines multi-omics data for cancer subtype classification. The presented framework utilises genomic, epigenomic, and proteomic data and combines them in a single model to enhance the accuracy of classification, while keeping the results biologically interpretable. Experimental results show that integrative, attention-based deep learning models, can learn molecular patterns associated with cancer heterogeneity and features that vary by subtype. Keeping that in mind, the list of contributions of this research are summarised as follows:

1. **Improved classification performance**: Testing results show that the presented framework reaches 92.3% of classification accuracy. The model performed better than the established baseline methods. It is concluded that deep learning models can learn complex, non-linear patterns in multi-omics, high-dimensional data.
2. **Effective multi-omics integration**: The presented model runs each omics data type through its own network branch. This mechanism gives a broad molecular view of cancer subtypes, and helps to study different associations across omics layers.
3. **Enhanced interpretability**: An attention mechanism was adopted to help the model point to the molecular features that most influence the classification decisions. The model's interpretability supports biological capability and helps identify biomarkers that distinguish between different subtypes.
4. **Robustness and generalisability**: Testing results show a good level of consistency when the data are disturbed, as it still predicts well on cancer types that were not part of the training. Results suggest that it can be generalised beyond the original associates.
5. **Subtype-specific biomarker identification**: Examining the attention weights of the presented model, it is concluded that molecular patterns linked to cancer subtypes can be picked out. These patterns can help in choosing candidate biomarkers and refining molecular classification.

Despite the advantages of the novel presented model, there are still some challenges to be mentioned here. For example, this model depends on large multi-omics data, which are not common in daily practice due to the high costs of collecting this kind of data.

Another limitation of this model, especially when it comes to multi-modal models that incorporate attention mechanisms, is that they are computationally costly, especially when there are limited computational resources. Even though this model has better interpretability compared to traditional deep learning methods, it does not completely address the black box problem of deep learning methods. There are issues of missing information and bias, especially when dealing with large public data, which can affect the final results of this model; thus, it is important to clean and validate the data before analysis.

This integrated deep learning framework shows a significant improvement in the classification of cancer subtypes through the integration of various omics modality data

types. This framework combines the benefits of high predictive accuracy with interpretability based on biological knowledge to identify the differences between various types of cancer and facilitate the process of data-driven approaches in cancer treatment. To use these approaches in the clinical setting, further refinement and validation of the accuracy of these approaches in well-designed studies are necessary.

It worth mentioning that this research has also identified several opportunities that could be further investigated in future research. These include the incorporation of clinical variables, medical imaging data, or single-cell omics data to improve prediction accuracy and better understand the biology. Other opportunities include further research on using transfer learning to extend the model to classify different types of cancer. This could improve the model's generalizability and minimize the need to label data in cases of rare cancer types. Further research could be conducted on developing techniques to better understand multi-omics deep learning models. These techniques could improve model interpretability while keeping in mind ethical considerations. To evaluate the efficiency of artificial intelligence-based cancer classification systems in real-world clinical settings and ensure that they meet ethical and public standards, research teams should establish channels of feedback from clinicians, patients, data scientists, and healthcare policymakers.

# References

[1] K. Cotter and M. A. Rubin, "The evolving landscape of prostate cancer somatic mutations," *The Prostate*, vol. 82, pp. S13–S24, 2022.

[2] M. O. Adebiyi, M. O. Arowolo, and O. Olugbara, "A genetic algorithm for prediction of rna-seq malaria vector gene expression data classification using svmkernels," *Bulletin of Electrical Engineering and Informatics*, vol.10, no.2, pp. 1071–1079, 2021.

[3] H. Aljawawdeh, A. Droubi, D. Mashaqbeh, and H. Alazzeh, "Transforming healthcare in Jordan: Enhancing the patient journey with AI innovations using
ai to enhance patient experiences and hospital efficiency," in *2024 Global Digital Health Knowledge Exchange & Empowerment Conference (gDigiHealth. KEE)*, pp. 1–8, IEEE, 2024.

[4] K. M. Jaber, R. Abdullah, and N. A. Rashid, "Hdt-hs: A hybrid decision tree/harmony search algorithm for biological datasets," in *2012 International Conference on Computer & Information Science (ICCIS)*, vol. 1, pp. 341–345, IEEE, 2012.

[5] T.Yue, Y.Wang, L.Zhang, C.Gu, H.Xue, W.Wang, Q.Lyu, andY.Dun, "Deep learning for genomics: A concise overview," *arXiv preprint arXiv:1802.00810*, 2018.

[6] K. Kourou, K. P. Exarchos, C. Papaloukas, P. Sakaloglou, T. Exarchos, and D. I. Fotiadis, "Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 5546–5555, 2021.

[7] O. Menyhárt and B. Győrffy, "Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis," *Computational and structural biotechnology journal*, vol. 19, pp. 949–960, 2021.

[8] C. Meng, O. A. Zeleznik, G. G. Thallinger, B. Kuster, A. M. Gholami, and A. C. Culhane, "Dimension reduction techniques for the integrative analysis of multi-omics data," *Briefings in bioinformatics*, vol. 17, no. 4, pp. 628–641, 2016.

[9] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, "Interpretability of machine learning-based prediction models in healthcare," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 5, p. e1379, 2020.

[10] J. Susan and P. Subashini, "Deep learning inpainting model on digital and medical images-a review.," *Internationa Arab Journal of Information Technology (IAJIT)*, vol. 20, no. 6, pp. 919–936, 2023.

[11] L. M"ullauer, "Molecular pathology of cancer: the past, the present, and the¨ future," 2021.

[12] C. M. Perou, T. Sørlie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, *et al.*, "Molecular portraits of human breast tumours," *nature*, vol. 406, no. 6797, pp. 747–752, 2000.

[13] F. Alharbi and A. Vakanski, "Machine learning methods for cancer classification using gene expression data: A review," *Bioengineering*, vol. 10, no. 2, p. 173, 2023.

[14] S. Mallik and Z. Zhao, "Graph-and rule-based learning algorithms: a comprehensive review of their applications for cancer type classification and prognosis using genomic data," *Briefings in bioinformatics*, vol. 21, no. 2, pp. 368–394, 2020.

[15] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[16] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning–based sequence model," *Nature methods*, vol. 12, no. 10, pp. 931–934, 2015.

[17] S. Yousefi, F. Amrollahi, M. Amgad, C. Dong, J. E. Lewis, C. Song, D. A. Gutman, S. H. Halani, J. E. Velazquez Vega, D. J. Brat, *et al.*, "Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models," *Scientific reports*, vol. 7, no. 1, pp. 1–11, 2017.

[18] J. Lanchantin, R. Singh, B. Wang, and Y. Qi, "Deep motif dashboard: visualizing and understanding genomic sequences using deep neural networks," in *Pacific symposium on biocomputing 2017*, pp. 254–265, World Scientific, 2017.

[19] R. Chen and M. Snyder, "Promise of personalized omics to precision medicine," *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, vol. 5, no. 1, pp. 73–82, 2013.

[20] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[21] T. Aittokallio, "Dealing with missing values in large-scale studies: microarray data imputation and beyond," *Briefings in bioinformatics*, vol. 11, no. 2, pp. 253–264, 2010.

[22] P. S. Reel, S. Reel, E. Pearson, E. Trucco, and E. Jefferson, "Using machine learning approaches for multi-omics data analysis: A review," *Biotechnology advances*, vol. 49, p. 107739, 2021.

[23] J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, and A. Telenti, "A primer on deep learning in genomics," *Nature genetics*, vol. 51, no. 1, pp. 12–18, 2019.

[24] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K.Ellrott, I.Shmulevich, C.Sander, andJ.M.Stuart, "Thecancergenomeatlas pan-cancer analysis project," *Nature genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.

[25] S. Andrews *et al.*, "Fastqc: a quality control tool for high throughput sequence data," 2010.

[26] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "Star: ultrafast universal rna-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.

[27] S. Anders, P. T. Pyl, and W. Huber, "Htseq—a python framework to work with high-throughput sequencing data," *bioinformatics*, vol. 31, no. 2, pp. 166–169, 2015.

[28] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, *et al.*, "The genome analysis toolkit: a mapreduce framework for

analyzing next-generation dna sequencing data," *Genome research*, vol. 20, no. 9, pp. 1297–1303, 2010.

[29] M. J. Aryee, A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen, and R. A. Irizarry, "Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays," *Bioinformatics*, vol. 30, no. 10, pp. 1363–1369, 2014.

[30] J. Li, Y. Lu, R. Akbani, Z. Ju, P. L. Roebuck, W. Liu, J.-Y. Yang, B. M. Broom, R. G. Verhaak, D. W. Kane, *et al.*, "Tcpa: a resource for cancer functional proteomics data," *Nature methods*, vol. 10, no. 11, pp. 1046–1047, 2013.

[31] S. Van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of statistical software*, vol. 45, pp. 1–67, 2011.

[32] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.

[33] A. Vaswani, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[35] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimisation framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2623–2631, 2019.

[36] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[37] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[38] R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, and O. Stegle, "Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets," *Molecular systems biology*, vol. 14, no. 6, p. e8124, 2018.

[39] Q. Mo, S. Wang, V. E. Seshan, A. B. Olshen, N. Schultz, C. Sander, R. S. Powers, M. Ladanyi, and R. Shen, "Pattern discovery and cancer gene identification
in integrated cancer genomic data," *Proceedings of the National Academy of Sciences*, vol. 110, no. 11, pp. 4245–4250, 2013.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[41] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz Jr, and K. W. Kinzler, "Cancer genome landscapes," *Science*, vol. 339, no. 6127, pp. 1546–1558, 2013.

[42] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[43] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genomewide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15545–15550, 2005.

[44] D. J. Slamon, B. Leyland-Jones, S. Shak, H. Fuchs, V. Paton, A. Bajamonde, T. Fleming, W. Eiermann, J. Wolter, M. Pegram, *et al.*, "Use of chemotherapy plus a monoclonal antibody against her2 for metastatic breast cancer that overexpresses her2," *New England journal of medicine*, vol. 344, no. 11, pp. 783– 792, 2001.

[45] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007.

[46] S. Huang, K. Chaudhary, and L. X. Garmire, "More is better: recent progress in multi-omics data integration methods," *Frontiers in genetics*, vol. 8, p. 84, 2017.

[47] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *cell*, vol. 144, no. 5, pp. 646–674, 2011.

[48] D. A. Fruman and C. Rommel, "Pi3k and cancer: lessons, challenges and opportunities," *Nature reviews Drug discovery*, vol. 13, no. 2, pp. 140–156, 2014.

[49] M. A. Gianfrancesco, S. Tamang, J. Yazdany, and G. Schmajuk, "Potential biases in machine learning algorithms using electronic health record data," *JAMA internal medicine*, vol. 178, no. 11, pp. 1544–1547, 2018.

[50] M. Schwaederle, M. Zhao, J. J. Lee, A. M. Eggermont, R. L. Schilsky, J. Mendelsohn, V. Lazar, and R. Kurzrock, "Impact of precision medicine in diverse cancers: a meta-analysis of phase ii clinical trials," *Journal of clinical oncology*, vol. 33, no. 32, pp. 3817–3825, 2015.

[51] N. Goossens, S. Nakagawa, X. Sun, and Y. Hoshida, "Cancer biomarker discovery and validation," *Translational cancer research*, vol. 4, no. 3, p. 256, 2015.

[52] S. Pushpakom, F. Iorio, P. A. Eyers, K. J. Escott, S. Hopper, A. Wells, A. Doig, T. Guilliams, J. Latimer, C. McNamee, *et al.*, "Drug repurposing: progress, challenges and recommendations," *Nature reviews Drug discovery*, vol. 18, no. 1, pp. 41–58, 2019.

[53] T. Choudhary, V. Mishra, A. Goswami, and J. Sarangapani, "A comprehensive survey on model compression and acceleration," *Artificial Intelligence Review*, vol. 53, pp. 5113–5155, 2020.

[54] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[55] W. N. Price and I. G. Cohen, "Privacy in the age of medical big data," *Nature medicine*, vol. 25, no. 1, pp. 37–43, 2019.

## Notes on contributors

***Hamzeh Aljawawdeh*** is a distinguished software engineering professional holding a PhD in the discipline. He has a solid background in academia and industry. He has significantly contributed to developing and advancing software engineering methodologies and practices. Hamzeh's research interests include business process models, software architecture, genetic algorithms, and e-learning, which have led him to produce numerous publications, conference presentations, and fruitful collaborations with other leading experts in the field. In addition, through their work, they have contributed to enhancing software engineering techniques and fostered a deeper understanding of the interplay between technology and educational practices. These accomplishments have cemented Hamzeh's reputation as a skilled and reliable software engineering professional.