

Pioneering AI Solutions for Cancer Subtype Classification through Gene Expression

Jayakrishnan Raveendran Pillai, Selvakumar Meera

Department of Computer Science and Engineering, Vels Institute of Science, Technology and Advanced Studies, Chennai, Tamil Nadu, India.

Abstract

Improved diagnostic models for personalized Cancer profiling are required significantly, utilizing AI methods to enhance accuracy, support early detection, and inform targeted treatment strategies. Despite significant progress in cancer prediction, current approaches often struggle with issues of generalizability across diverse patient cohorts, computational inefficiencies, and managing heterogeneous data sources. This paper delves into the fast developing topic of AI-driven tumor class categorization utilizing expression of genes data. Focusing on machine learning (ML), explainable artificial intelligence (XAI), neural network, and transfer learning techniques. The integration of innovative AI methodologies is crucial for understanding complex genetic interactions, improving model interpretability through XAI, and enabling adaptive learning through transfer learning. This will allow medical practitioners to rely on AI-driven insights and provide strong, scalable solutions for everyday life applications in medicine. The analysis recognizes existing limitations, including the absence of established methods on cross-institutional sharing of information and the difficulties in maintaining model adaptation to different tumor subtypes. This work underscores the potential of AI to revolutionize cancer subtype classification, fostering advancements that could reshape personalized oncology, improve patient outcomes, and establish a new standard for precision medicine. Unlike prior reviews, this study goes beyond summarizing methods by synthesizing cross-cutting gaps across ML, neural network (NN), XAI, and transfer learning (TL) approaches. It further proposes a conceptual framework that integrates these methodologies to guide future research in developing clinically deployable and patient-centered cancer diagnostic systems.

Keywords: Cancer profiling- AI-driven cancer subtype classification- gene expression- machine learning- explainable AI

Asian Pac J Cancer Biol, **10** (4), 1043-1060

Submission Date: 07/24/2025 Acceptance Date: 09/16/2025

1. Introduction

The disease is also the second leading cause of death worldwide, resulting from aberrant cell development and metastatic growth [1]. The cells of cancer frequently multiply independently of development signals and do not respond to survival/death signals, resulting in apoptosis. This phenomenon is caused by inherited factors, such as DNA mutations or epigenetic modifications. Some cancer genes, like BRCA1/2, are inherited that possess a strong depth because of their involvement in cellular control. [2]. Analysing unregulated gene transcription systems in cancer cells could aid in early diagnosis and treatment. Identifying certain genes (gene signatures) can lead to

an accurate diagnosis as well as more targeted therapy choices. Microarray analysis and RNA-seq devices allowed researchers to develop and evaluate novel mathematical and statistical models to evaluate genetic expression data, Calculating transcript concentrations across thousands of domains over a wide range of human patient samples [3]. Express technology has transformed the study of gene expression by allowing simultaneous assessments of gene alterations under a variety of experimental circumstances. This has allowed for the identification of disease genes, therapeutic targets, and tumour subtypes [4, 5]. Some significant genes are connected with particular cancer

Corresponding Author:

Dr. Selvakumar Meera

Department of Computer Science and Engineering, Vels Institute of Science, Technology and Advanced Studies, Chennai, Tamil Nadu, India.

Email: meeraselvakumar@gmail.com

subtype classifications that can be submitted to the FDA for validation and diagnosis [6]. Furthermore, the affected gene space frequently comprises noisy and redundant genes, which might have a negative impact on classification performance. As an illustration, the k-nearest-neighbour technique was prone to useless categorization properties [7].

Data mining techniques are typically classified into three categories: unsupervised learning, supervised learning, and reinforcement learning. Supervised learning uses a labelled training collection for mapping the input information into the proper output [8]. Unsupervised learning, in contrast, is not dependent on designated data; instead, the approach discovers information and pattern architectures. Such as clustering, on its own. In this context, the model's role is to detect patterns or group the input data into meaningful classes. Supervised learning typically involves a classification task, where the goal is to assign data to predefined categories. In unsupervised learning, clustering is a typical method used to explore the underlying data distribution, often serving as a pre-processing step for feature selection [9].

Since gene expression data has increased dramatically, various methods for analysing and diagnosing disease utilizing ML techniques have been developed. Using gene expression data analysis, these approaches classify samples according to their anticipated survival status. Methods based on ML are currently developed for analysing expression profiles. However, the elevated dimensionality in data from microarray gathering, along with limited sample sizes, restricts statistical power for clinical applications [10]. this frequently results in overfitting of pattern profiles, resulting in poor generalization capability [11]. Traditional ML algorithms, such as Cox's proportional hazard model along with encouragement vector machines, are frequently used for forecasting and recognizing cancer [12]. Deep learning (DL) models and algorithms are currently receiving a lot of interest from scientists and researchers throughout the world. DL, a subset on ML, takes advantage of advances in neural network technology. It functions by incorporating multiple hidden layers, activation functions, and hyper parameter tuning to process inputs and generate outputs. This structure makes DL models more sophisticated and advanced, offering substantial advantages in classification tasks. They are very good at handling complex and huge datasets, outperforming typical predictive models. For the past few years, DL has contributed to significant advances in healthcare, particularly in health imaging and cancer diagnosis.

This paper primarily examines the latest advancements in ML and DL techniques in classification of cancer. The increasing availability of healthcare data, along with the advancement of data analysis tools, has significantly improved the use of ML and DL in the healthcare sector [13]. Both ML and DL have made remarkable progress in addressing various scientific challenges [14]. In medical care, the use of AI serves a major part in several applications, including data management. Drug development, disease forecasting, and treatment planning

[15].

- This research examines mechanical learning as well as deep neural network models towards cancer subtype categorization, including their techniques, strengths, and limitations in processing gene expression data.

- It highlights the role of multi-omics data integration, including RNA-Seq and ATAC-Seq, to enhance diagnostic accuracy and subtype classification, while discussing how these approaches address challenges related to tumor heterogeneity and complex data patterns.

- The review outlines critical limitations such as computational demands, generalizability issues, and data inconsistencies, offering targeted recommendations for future research, including standardized data protocols, optimized algorithms, and resource-efficient deployment.

- This paper proposes strategies to improve model interpretability and scalability for clinical settings, emphasizing the potential of AI techniques to bridge gaps in clinical usability and patient-centred diagnostic outcomes

2. Background of the Study

The research on cancer subtype classification highlights the revolutionary impact. Of AI and ML in cancer diagnostics, especially through the analysis of gene expression data shown in Figure 1 [16]. Traditional ML models, like support vector machines (SVMs) and decision trees, initially offered moderate success. DL models like CNN and RNN have advanced the field by capturing intricate patterns within complex gene expression datasets. XAI has made AI-driven predictions more interpretable for clinicians, addressing transparency in medical decision-making. However, challenges remain, such as data heterogeneity, high computational costs, and model generalizability. Recent research has turned to strategies like transfer learning, multi-omics integration, and enhanced data pre-processing techniques for more robust, scalable, and clinically applicable AI models.

Nethala et al. [17] proposed the Optimal Gene Therapy Network (OGT-Net), an advanced AI-driven framework for classifying various types of cancers using gene expression data. The method integrates dataset normalization, feature extraction through Light Gradient Boosting Model (LGBM), and optimal feature selection using Interrupt-based Harris Hawk Optimization (IHHO)

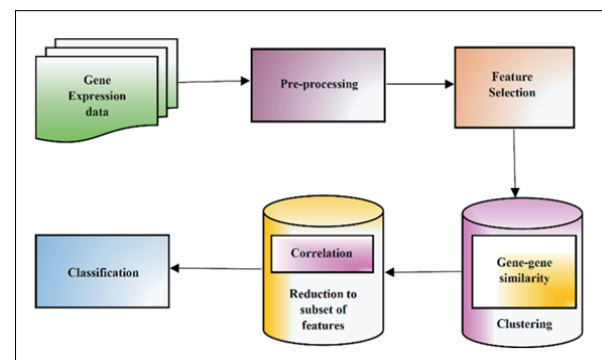


Figure 1. Cancer Classification Model Using Gene Expression Data

to remove redundant gene sequences. Subsequently, a customized DL convolutional neural network (DLCNN) is employed to categorize cancers including lymphography, colon, lung, ovarian, and prostate types. Simulation results demonstrate that OGT-Net outperforms state-of-the-art approaches, achieving an average accuracy of 91.13%, precision of 90.84%, recall of 91.25%, and F1-score of 90.7%, reflecting significant improvements over existing methods. The framework emphasizes both performance enhancement and clinical applicability, highlighting the potential for integration into user-friendly interfaces for healthcare practitioners. While OGT-Net shows promise in robust cancer classification, future research could focus on optimizing model architecture, improving interpretability, and bridging the gap between computational advancements and practical clinical deployment.

Abidalkareem et al. [18] developed a ML-based framework for identifying stage-specific biomarkers in breast cancer using dysregulated microRNAs (miRNAs). Leveraging a dataset of 1097 metastatic tissue samples from TCGA, the study applied Neighborhood Component Analysis (NCA) and Minimum Redundancy Maximum Relevance (MRMR) for feature selection to isolate the most discriminant up- and down-regulated miRNAs across the four stages of breast cancer. Both methods significantly outperformed the conventional fold-change (FC) approach, with NCA achieving the highest classification accuracy of 98.3% and MRMR reaching 93.1%. While NCA proved effective in identifying stage-specific biomarkers, MRMR provided complementary information by highlighting common biomarkers relevant across multiple stages. The study underscores the potential of advanced feature selection in improving diagnostic precision and facilitating early detection, though a key limitation remains the inability to incorporate blood samples, which share similar miRNA profiles in normal and cancerous tissues.

Babichev et al. [19] investigated the application of DL architectures for cancer classification using gene expression data, comparing CNNs, LSTMs, GRUs, and hybrid models. To optimize performance, the authors employed Bayesian optimization with 5-fold cross-validation for hyperparameter tuning and introduced a hybrid quality criterion, integrating an F1-score through the Harrington desirability method. Their framework follows a hierarchical step-by-step processing approach, where predictions from individual DL models are refined through a CART-based classifier to enhance decision-making objectivity. Experimental evaluation on datasets covering eight cancer types and a normal sample subset revealed that a two-layer GRU-RNN achieved the highest performance, with an accuracy of 97.8%. The study highlights the robustness of GRU-based recurrent networks for gene expression classification, while also emphasizing that increasing model complexity through ensembles may not guarantee superior accuracy. Nevertheless, the incorporation of hybrid decision-making mechanisms offers improved interpretability and reliability.

Al-Azani et al. [20] addressed two key challenges in gene expression-based cancer classification: class imbalance and the curse of dimensionality. The study applied oversampling techniques, including SMOTE and its variants, to balance the datasets at the data level, while ensemble learning was adopted at the algorithmic level to improve robustness. To reduce dimensionality, chi-square and information gain methods were first applied independently, and then combined into a novel hybrid feature selection approach (CHiS-IG) to identify the most informative genes. Among the evaluated models, the integration of SVM-SMOTE with a random forest (RF) classifier achieved the best performance, reaching 100% accuracy in some datasets, surpassing results reported in prior literature. The findings highlight the effectiveness of combining oversampling and hybrid feature selection in mitigating the limitations of high-dimensional, imbalanced gene expression data.

Khalsan et al [21], created a new fuzzy gene selection technique (FGS) enhance tumor categorization using expression of genes data. The approach uses three methods for selecting features (Mutual Information, F-ClassIf, and Chi-squared) to identify relevant genes while reducing the dimensionality of the data. The blurring and the defuzzification methods were used to identify the highest overall rating for each gene, which assisted in identifying essential genes. The FGS-enhanced cancer classification model outperformed the classic MLP method with regard to of accuracy, precision, recall, and the f1-score. (96.5%, 96.2%, 96%, and 95.9%, respectively). The suggested model exhibited its ability to accurately classify cancer in six datasets, indicating its promise in a variety of domains, including biomedical science. However, the FGS model's high computational demands may limit its scalability and efficiency with large gene expression datasets. The fuzzy gene selection strategy reduces the curse of dimensionality by filtering irrelevant genes, thereby improving robustness in handling high-dimensional expression data.

Joshi et al [22], researched introduces rPAC, a new pathway analysis paradigm that divides signalling pathways are divided under two parts: The precursor component of a transcriptional element blocking includes downstream section piece that makes up a TF block. The grading method is then used on a collection on expressed genetic material data sets, resulting in the following summary metrics: "Proportion of Significance" (PS) and "Average Route Score" (ARS). The method's performance was evaluated using both simulated data and an actual investigation comprising three epithelial tumor datasets using the Cancer Genome Atlas (TCGA). The rPAC approach highlighted various pathways as potential forms of cancer endorsements, and it was discovered to be more effective than standard methods in detecting illness etiology, especially when distinguishing pathways and sections of damaged circuits at a greater resolution.. However, the rPAC method's complexity may hinder its scalability and adaptability to diverse gene expression datasets and pathway structures. By decomposing signaling pathways into modular components, rPAC captures hierarchical biological relationships, improving

interpretability of gene expression patterns across different cancer subtypes.

Nahiduzzaman et al [23] described a new method for reliably classifying three forms of lung cancer, as well as normal lung tissue, using CT images. The method makes use of a compact parallel depth-wise separable CNN (LPDCNN) with the ridge regress radical training device. The method improves image quality and decreases noise by employing contrast-limited adaptive histogram equalization (CLAHE) and Gaussian blur. The LPDCNN retrieves discriminant characteristics with little computational cost. This Ridge-ELM approach was developed for better the accuracy of classification. The structure achieves average recall and accuracy values of $98.25 \pm 1.031\%$ and $98.40 \pm 0.822\%$ in four-class categorization, respectively. It is also extremely efficient, with testing times of only 0.003 seconds. The system also includes SHAP (Shapley Additive Explanations) to improve explain ability and decision-making in real-world lung cancer. However, the method's dependence on specific pre-processing techniques may restrict its generalizability to diverse imaging conditions and lung cancer subtypes. The use of depth-wise separable convolutions significantly reduces computational cost while preserving discriminative features, making it more efficient for large-scale imaging and gene-linked diagnostic tasks.

The article examines recent developments in ML algorithms for categorized cancer and subtype detection, focusing on DL frameworks like adversarial networks and convolutional architectures. These methods enhance the detection of tumor origins, molecular subtypes, and gene interaction groups. However, challenges remain in scalability, computational efficiency, and adaptability across diverse datasets and cancer types. Addressing these limitations is essential because moving the models beyond studies towards clinical use, thereby improving cancer diagnosis and treatment planning across a variety of biological and imaging contexts.

3. Review Analysis

The review paper on tumor category classifying utilizing genetic expression data provides a complete methodology. That combines ML, explainable AI classifiers, neural network architectures, and transfer learning to improve diagnostic accuracy and clinical relevance. ML techniques effectively analyze complex gene expression profiles to identify distinct cancer subtypes, while explainable AI enhances interpretability, allowing for insights into the biological significance of predictive markers. Neural network architectures, particularly DL models, capture intricate biological interactions and facilitate the identification of subtype-specific biomarkers the technique of transfer learning uses models that have been trained to improve classification performance, particularly in settings with little data. Moving into the future, subsequent research must concentrate on integrating multi-omics data to further refine classification accuracy, enhancing model robustness against diverse datasets, and developing real-time

applications for clinical decision-making, ultimately advancing personalized cancer treatment strategies.

From the comparative review, three cross-cutting gaps emerge: (i) data-related challenges, such as heterogeneity, small sample sizes, and lack of standardized protocols; (ii) model-related challenges, including overfitting, computational inefficiency, and poor scalability; and (iii) clinical translation gaps, such as limited interpretability, lack of cross-institutional validation, and barriers to clinical adoption. Addressing these gaps requires a unified approach that combines predictive accuracy, interpretability, and adaptability.

3.1 ML based on approaches

A ML approach for classifying Tumor types depend on information on gene expression begins with pre-processing steps such as normalization and gene filtering to reduce noise as shown in Figure 2 [24]. Feature selection methods like t-tests or Recursive Feature Elimination (RFE) help identify the most relevant genes, while dimensionality reduction techniques like PCA or auto encoders reduce the feature space while retaining critical information. Classification is then performed using a variety of models, such as SVM, RF, and DL structures. The efficacy of the model is measured utilizing metrics such as precision as well as ROC-AUC, with cross-validation to ensure robustness and generalizability.

Rukhsar et al [25], provided a novel strategy for dealing with Multidimensional and loud. RNA-Seq data from the Mendeley repository, with the goal of retrieving information on five forms of cancer. Eight DL algorithms are used to pre-process the data, extract features, and classify the results. CNN outperformed all other algorithms, and the study classified five tumours using related genes. Comparative research revealed that the proposed technique outperformed the existing literature. However, High computational demands and risk of overfitting may limit the approach's scalability and generalizability for RNA-Seq data.

Sun et al [26], proposed SCM-DNN, to find specific the simultaneous expression circuits for each molecular category, enabling for better and more precise predictions for breast while stomach cancer patients. SCM-DNN beats standard gene expression-based approaches across all criteria, even with imbalanced sample sizes. The discovered genes may represent particular subtype features and help to improve knowledge about molecular subtyping processes, potentially driving personalized therapy. The incorporation of multi-omics data is

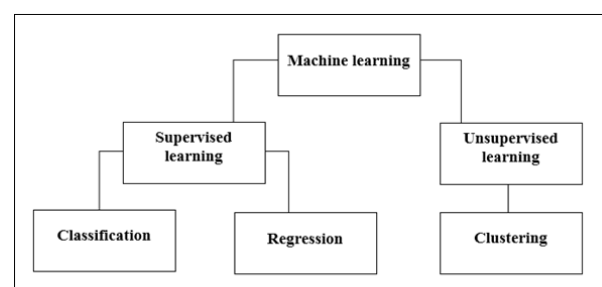


Figure 2. Types of Classification in ML

recognized being a useful technique for studying biological systems. Despite, A potential limitation of SCM-DNN is its reliance on high-quality, large-scale multi-omics data, which not always be available, potentially affecting model accuracy and generalizability.

Park et al [27], RNA-seq transcriptome and ATAC-seq epigenetic data integrated to create a system of classification for cancer of the breast fundamental categories. It identifies eleven important genes involved in immunological comments, hormone signalling, progression of cancer, and cell division. The study employs bulkRNA-seq and ATAC-seq data to investigate the connection between the expression of genes and access to chromatin in cancer of the breast patients. The research shows that integrating RNA-seq and ATAC-seq data with ML algorithms may improve the comprehension of the accessibility of chromatin and the molecular mechanisms that drive these subtypes. However, a limitation of the study is the potential bias in the ML model due to the reliance on bulk RNA-seq and ATAC-seq data, which not capture the heterogeneity of tumour cell populations.

Babichev et al. [28] proposed a hybrid ML framework to evaluate proximity metrics for high-dimensional gene expression data, focusing on their role in clustering and disease classification. By integrating data mining methods with ML techniques such as k-medoid clustering, RF, Bayesian optimization, and a stacking meta-classifier, the model achieved high accuracy (>95.9%) across 13 TCGA cancer types and demonstrated strong generalizability on Alzheimer's and Type 2 Diabetes datasets. A key contribution was the comparative analysis of correlation distance, mutual information, and Wasserstein metrics, with correlation and Wasserstein proving highly effective and interchangeable for clustering and classification. The stacking model further enhanced robustness against clustering errors, enabling a scalable and automated pipeline suitable for precision medicine. This work highlights how metric-driven hybrid modeling can support reliable biomarker discovery and early disease diagnostics from gene expression profiles.

Alanazi et al. [29] proposed an integrative ML framework for classifying cancer subtypes using RNA-seq data from BRCA, KIRC, COAD, LUAD, and PRAD. The approach combined data normalization, feature selection, dimensionality reduction, clustering (k-means), and classification, ultimately employing a Wide Neural Network that achieved remarkably high accuracy (99.995% on the test set). The method was proposed to address the limitations of traditional histopathology by leveraging transcriptomic signatures for more precise and personalized cancer subtype stratification. The key benefit lies in its ability to unravel molecular heterogeneity and improve diagnostic accuracy, paving the way for precision oncology and tailored therapies. However, the authors note that the reliance on large, high-quality RNA-seq datasets poses a limitation, as data biases or noise could reduce robustness and affect the reliability of clinical applications.

Babichev et al [30], introduced a hybrid inductive approach to creating uniquely expressed and socially connected expression patterns using the technique of

spectral clustering. The model proved to be insufficient for internal as well as external quality standards, leading the creation of a balancing clustering quality criterion. The best cluster structures were discovered to be four and six cluster configurations. The algorithm's appropriateness has been evaluated using a classifier on gene expression datasets. The method of RF as the CNN were utilized to handle binary classification and multiclass identification issues. However, the model's reliance on specific clustering configurations may limit flexibility in identifying gene expression patterns across diverse datasets.

Liu et al [31], proposed ML-based method has been developed to generate a consensus immune-related lncRNA signature (IRLS), this constitutes an individual contributor to risk for survival in general. IRLS provides consistent results but has a modest value in predicting for relapse-free survival. It can be more accurate with conventional diagnostic and molecular features. The group with a higher risk is more susceptible to fluorouracil-based chemotherapy, whereas the group with a low risk benefits better from bevacizumab. IRLS may improve medical results among particular patients suffering from CRC. However, the low estimated value for relapse-free mortality may limit the signature's overall utility in clinical decision-making.

Mohamed et al [32], researched developed a hybrid approach to cancer of the breast. Identification and diagnosis that integrates the Ebola optimization search algorithm (EOSA) Using a CNN design that utilizes expression of genes data. The data was pre-processed using a variety of approaches, including outlier removal, normalization, filtering, and conversion to two-dimensional pictures. The previously EOSA-CNN technique was used in categorization. The predictive model performed better than the malignant category with respect to of accuracy, precision, recall, f1-score, kappa, preciseness, and sensitivity. The findings suggest that the model may accurately and consistently diagnose breast cancer utilizing genetic expression data. Future improvements will address unbalanced data and integrate the model with new optimization algorithms. However, the model's performance may be compromised by the challenges of handling imbalanced data, which could affect classification accuracy.

Sarkar et al [33], proposes a genetic algorithm and SVM (GA-SVM) breast cancer classification model that employs a combination of ML methodologies. Using clinical pathology data from numerous tertiary care hospitals, the model differentiates between those suffering from a triple-negative and non-triple-negative breast disease. Whenever utilized with two independent medical facilities datasets from the North West Africa peninsula, the model outperformed the other models. A ten-fold cross-validation handle was employed to guarantee that the framework utilized for prediction accuracy remained consistent across all models. The model's efficacy was assessed using measures such as average square mistake, logarithm loss, F1 value, ROC curve, and precision-recall graph. However, the model's reliance on specific clinical pathological data may limit its generalizability across

different populations and cancer types.

Senbagamalar and Logeswari [34] addressed the challenge of multiclass cancer classification using gene expression data by proposing a genetic clustering algorithm (GCA) for optimal feature selection and a divergent RF (DF) classifier. Their approach reduced 1621 gene features to just 21 highly informative ones, enabling efficient classification of five cancer types: breast, colon, kidney, lung, and prostate cancer. The proposed GCA-DF model achieved 95.21% accuracy, 93% specificity, and 94.29% sensitivity, outperforming conventional classifiers. By combining clustering-based feature reduction with an ensemble classifier, the study highlighted the importance of compact yet discriminative gene subsets in improving diagnostic accuracy. The authors further suggested incorporating metaheuristic optimization strategies in the future to refine gene expression selection and enhance computational efficiency in large-scale cancer diagnostics (Table 1).

Recent advances in cancer research have used ML approaches to increase diagnostic and prognosis accuracy. Techniques such as RNA-Seq data processing and SCM-DNN models have improved cancer subtype prediction and noise reduction, but they have computational and generalizability limitations. Combining RNA-Seq and ATAC-Seq data improves comprehension of chromatin accessibility, but it may neglect tumour heterogeneity. Models such as AWCA and LGDLDA are highly concordant with existing classification standards, but their applicability to varied populations remains a challenge. The Consensus Immune-Related lncRNA Signature (IRLS) has improved colorectal cancer prognosis, while approaches such as EOSA and GA-SVM show promise in breast cancer diagnosis. The Knowledge- and Context-Driven ML (KCML) paradigm demonstrates the power of ML in large-scale genetic studies.

3.2 Explainable AI Classifier based on approaches

Explainable AI (XAI) classifiers are used to categorize Tumor classifications based on expression of genes information. As shown in Figure 3 [35]. These classifiers increase interpretability and predictive performance, allowing researchers and clinicians to better understand the decision-making process involved in categorization. XAI classifiers use techniques Examples include decision tree

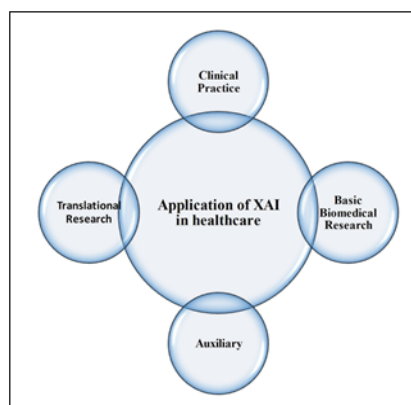


Figure 3. Application of XAI in Healthcare

structures, randomly generated forests, and SVM, among others, to highlight essential features and interactions in gene expression data, confirming predictions and building trust among medical practitioners. This transparency helps to uncover biomarkers linked with various subtypes and facilitates individualized treatment regimens, which ultimately improves patient outcomes. However, obstacles like data quality, dimensionality, and biological system complexity persist, demanding additional study for greater clinical application.

Wani et al [36], introduced “DeepXplainer,” a hybrid DL system that detects lung cancer and explains forecasts. It makes use of a CNN and XGBoost to predict class labels. The technique uses an understandable computational intelligence method known as “SHAP” to provide explanations. This algorithm was used to process the freely available “Survey Lung Cancer” information and exceeded previous methods with respect to precision, responsiveness, and F1 score. The model had 97.43% accuracy, 98.71% sensitivity, and an F1-score of 98.08. Each forecast is provided using an intelligible AI method on both a local and global level. Data diversity across different lung cancer subtypes and patient populations may provide a challenge.

Li et al [37], proposed a new framework (CGMega) has been created to analyze cancer gene modules using explainable graph attention. The system uses a multi-omics representation chart, with vertices representing genomes and lines denoting interactions between proteins. It detects cancer-related genes using a transformer-based graph attention neural network within a semi-supervised setting. Further outstanding performance of CGMega enables the subsequent detection of cancer gene modules. GNNExplainer40 is a model-agnostic method for understanding contributing variables to cancer genes in a multi-omics setting. CGMega was tested on cancerous breast cell lines and AML patients, finding high-order linkages among gene in disease genetic networks. However, May face limitations in generalizability across diverse cancer types and multi-omics datasets.

Abhang and Gunjal [38] proposed a Deep Graph Ensemble CNN (G-ECN) for drug response prediction using multi-omics cancer cell line data from the GDSC2 dataset. The model integrates graph-based multi-scale feature representation with transfer learning from gene ontology knowledge to capture critical gene–drug interactions, addressing the challenge of accurately distinguishing “sensitive” and “resistant” drug responses. This approach was proposed to improve personalized cancer therapy by leveraging structural and genomic features beyond traditional models. The main benefits include higher predictive accuracy, strong generalizability across datasets (validated on CCLE), and explainable outputs aligned with biological knowledge, making it useful for precision oncology. However, the method relies heavily on high-quality multi-omics data and remains computationally expensive, which may limit scalability in clinical settings.

Sekaran et al [39], described OPSCC is an unpredictable disease having a poor prognosis. That significant medical

Table 1. Comparison of Literature Done in ML Based on Approaches

Author/Reference	Technique	Significance	Limitation
Rukhsar et al [25]	RNA-Seq data	- Reduces noise in RNA-Seq data - Enabling accurate cancer classification.	High computational demands and risk of overfitting may limit the approach's scalability and generalizability for RNA-Seq data.
Sun et al [26]	SCM-DNN	- Enhances cancer subtype prediction - Supporting personalized therapies.	Potentially affecting model accuracy and generalizability.
Park et al [27]	RNA-seq transcriptome and ATAC-seq epigenetic data	- Enhances understanding of chromatin accessibility - Facilitating improved classification of intrinsic subtypes.	Which not capture the heterogeneity of tumour cell populations.
Babichev et al. [28]	Hybrid ML framework with proximity metrics (k-medoid clustering, Random Forest, Bayesian optimization)	- Achieved >95.9% accuracy across 13 TCGA cancers. - Demonstrated strong generalizability to Alzheimer's and Type 2 Diabetes.	Relies heavily on metric selection; performance may vary with dataset characteristics and metric suitability.
Alanazi et al. [29]	Integrative ML pipeline with normalization, feature selection, dimensionality reduction, clustering (k-means), and Wide Neural Network	- Achieved exceptionally high accuracy (99.995%) on RNA-seq cancer subtype data. - Addressed histopathology I imitations by leveraging transcriptomic signatures.	Requires large, high-quality RNA-seq datasets; data noise or bias can reduce robustness and clinical applicability.
Babichev et al [30]	Hybrid inductive model	- Enhances the identification of distinct gene expression profiles - More accurate classification in gene expression analysis.	Limit flexibility in identifying gene expression patterns across diverse datasets.
Liu et al [31]	Consensus IRLS	- Enhances prognostic accuracy for colorectal cancer - Improve patient outcomes	Limit the signature's overall utility in clinical decision-making.
Mohamed et al [32]	EOSA	Improves lung cancer identification with genetic activity information.	Model performance may be compromised by the challenges of handling imbalanced data
Sarkar et al [33]	GA-SVM	GA-SVM model enhances breast cancer classification accuracy	Limit its generalizability across different populations and cancer types.
Senbagamalar & Logeswari [34]	GCA + DF	- Reduced 1621 genes to 21 discriminative features. - Achieved 95.21% accuracy, 93% specificity, and 94.29% sensitivity.	Future improvement needed via metaheuristic optimization for better scalability and computational efficiency.

conditions along with elevated recurrence rates are connected to current therapies. Highlight the importance of advancing diagnostic techniques in OPSCC. Researching biomarkers that are especially those that may be gathered without intrusive procedures, has the potential to alter patient care methods. The current study used a genomic approach to uncover oncogenic factors involved in the development of OPSCC. Combining informatics studies and ML methodologies with a detailed examination the RNAseq information resulted in the discovery of ECT2, LAMC2, and DSG2 as potential molecular markers for OPSCC. The study's findings may help improve the survival rates of OPSCC patients. The findings and technique of this study could be used to clinical and experimental settings in future research. However, the findings may lack generalizability due to potential biases in the RNA-seq datasets used for analysis.

Morabito et al [40], the article presents the DeepSHAP Auto encoder Filter for Gene Selection (DSAF-GS), a new DL and XAI-based FS approach for genomics-scale data analysis. The technique uses AEs to select those that are most useful genes while retaining the initial characteristic time, boosting the explain ability of results and using AEs' representation capabilities. Gene selection is used to build

and train diagnostic or prognostic prediction models. The Shapely Additive Ex-Planation (SHAP) XAI technique is then used to examine the model findings and determine the genes that are most related to the condition. A systematic population of newly identified Bi net stage A CLL patients was studied using the XAI approach to discover markers which levels of expression predict the need for medication. Despite, May have limited scalability with large, complex genomic datasets due to computational demands.

Yang et al. [41] introduced MHGCN, a Multi-channel Hybrid Graph CNN, for cancer drug response prediction by explicitly modeling the topological relationships of cell line–drug pairs (CDPs). The framework integrates gene expression data and drug molecular fingerprints, refines CDPs with denoising autoencoders, and builds both a similarity network (via cosine similarity) and a heterogeneous response graph. MHGCN jointly learns from these graphs using graph convolutional layers and fuses embeddings through a weighted matrix projection to generate predictions. This model was proposed to overcome the limitation of prior methods that ignored intrinsic CDP connections. The main advantage lies in its improved predictive accuracy and ability to capture complex biological interactions, making it valuable for

Table 2. Comparison of Literature Done in Explainable AI Classifier based on Approaches

Author/Reference	Technique	Significance	Limitation
Wani et al [36]	DeepXplainer	· Improves lung cancer prediction accuracy with clear, interpretable explanations.	Limitations in handling data variability across diverse lung cancer subtypes and patient populations.
Li et al [37]	CGMega	· Enhances detection of cancer gene modules	Faced limitations in generalizability across diverse cancer types and multi-omics datasets.
Abhang & Gunjal [38]	G-ECN	· Improves drug response prediction with high accuracy and biological interpretability.	Requires high-quality data; computationally expensive.
Sekaran et al [39]	OPSCC	· It aims to improve patient outcomes and survival rates through early detection and targeted therapies.	Lack generalizability due to potential biases in the RNA-seq datasets used for analysis.
Morabito et al [40]	DeepSHAP	· enhances diagnostic and prognostic gene selection accuracy	Limited scalability with large, complex genomic datasets due to computational demands.
Yang et al. [41]	MHGCN	· Captures complex CDP interactions and enhances predictive accuracy.	Dependent on dataset quality; high computational cost.
Gutierrez-Chakraborty et al [42]	XAI framework	· Demonstrating high predictive and therapeutic relevance	Framework's applicability may be limited by its focus on a specific demographic
Abuzinadah et al [43]	Predictive model	· Enhances cancer prediction reliability · Achieving 96.87% accuracy	Leads its generalizability to other cancer types or diverse data sources.
Altini et al [44]	CAD	· Improves breast tumour assessment by combining accuracy with interpretability · SHAP-based analysis enhances understanding of model decisions	Model may limit its effectiveness across varied clinical scenarios.
Rajpal et al [45]	XAI-CNVMarker	· AI-powered platform for identifying interpretable biomarkers in breast cancer. · The model had a classification accuracy of 0.712 with a 95% confidence interval.	Reliance on specific datasets for validation could restrict the generalizability of the identified biomarkers across diverse patient populations.

personalized therapy. However, its performance depends on the quality and completeness of high-throughput datasets, and the computational cost of multi-channel graph learning may restrict scalability in real-world clinical environments.

Gutierrez-Chakraborty et al [42], researched an XAI framework for artificial intelligence methodology for finding and validating essential genetic indicators for HCC prediction. The technique involves evaluating medical and data on gene expression to discover possible biomarkers with predictive value. The research employs advanced artificial intelligence algorithms that have been established against large genetic expression datasets, proving the biomarkers' accuracy in predicting and therapeutic utility. Key biomarkers such as TOP3B, SSBP3, and COX7A2L have been shown to be influential in many models, increasing HCC prognosis beyond AFP. These biomarkers are also relevant to the Hispanic community, which is consistent with the overall purpose of demographic-specific research. However, the framework's applicability may be limited by its focus on a specific demographic, potentially hindering generalizability to other populations.

Abuzinadah et al [43], proposed predictive model employs a stacking ensemble approach that combines the advantages objective both boosting and bagging classifiers, with the objective of increasing prediction accuracy and reliability. This combination minimizes variation while improving generality, yielding better cancer forecasting findings. The suggested approach achieves 96.87% accuracy, this represents the greatest performance of the model recorded on this set of data to date when every attribute are considered. The data is evaluated with SHAPly, which is an explainable artificial intelligence technique. When contrasted with other

cutting-edge models, the suggested model outperforms them. Despite, the model may face limitations in scalability and adaptability when applied to different datasets or cancer types, potentially impacting its broader applicability.

Altini et al [44], created an explainable computer-aided diagnosis (CAD) system to help pathologists assess tumour cellularity in breast histopathology slides. The system compared an end-to-end DL technique that used a Mask R-CNN segmentation instance architecture to a two-stage procedure that extracts features based on the morphology and textured properties of cell nuclei. SVM algorithms and ANN are used to develop classifiers that can differentiate among neoplastic and non-tumor nuclei. Overall SHAP explainable artificial intelligence method has been applied to evaluate feature significance, providing a clearer understanding of the judgments made by ML models. An experienced pathologist validated the model to ensure its clinical usefulness. However, two-stage pipeline models are substantially less precise, they are easier to interpret, which may increase confidence in using AI-based CAD systems in clinical processes. However, The CAD system's reduced accuracy in the two-stage model may limit its effectiveness across varied clinical scenarios.

Rajpal et al [45], researched XAI-CNVMarker, an AI-powered platform for identifying interpretable biomarkers in breast cancer. The approach use DL to create a classification model, which is then examined using explainable AI techniques to find 44 CNV biomarkers. The model had a classification accuracy of 0.712 with a 95% confidence interval. The biomarkers were also validated using METABRIC, illustrating the significance of transparent artificial intelligence in identifying practical

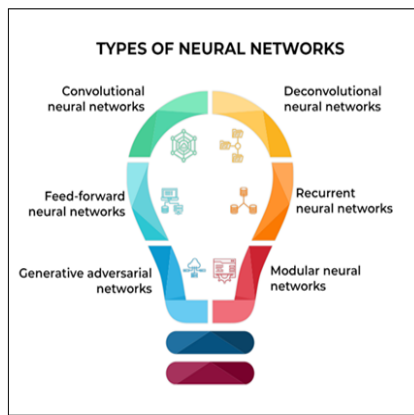


Figure 4. Type of Neural Network Architectures

indicators. However, reliance on specific datasets for validation could restrict the generalizability of the identified biomarkers across diverse patient populations (Table 2).

Recent developments in cancer diagnosis have used AI and DL to improve forecast accuracy and interpretability. DeepXplainer and CGMega are models that improve lung cancer forecasts, but their generalizability across cancer types is limited. Deep GONet is highly accurate across gene expression datasets, although it suffers with novel gene connections. The OPSCC model tries to improve patient outcomes by early identification, however it may be influenced by RNA-seq data. DeepSHAP enhances gene selection accuracy but has scalability issues with huge datasets. PathDeep improves cancer biology understanding for tailored medicines, however it may struggle with limited data. The XAI methodology has strong predictive value, however it may not generalize well to different cancer types.

3.3 Neural network based on approaches

Neural networks have emerged as powerful tools for cancer subtype classification due to their ability to model complex, non-linear relationships in high-dimensional gene expression data. CNNs are particularly effective at extracting local co-expression patterns among groups of genes, while recurrent neural networks (RNNs) capture sequential dependencies in gene regulatory pathways. Autoencoders and deep embedding frameworks reduce dimensionality and denoise input data, enabling the extraction of biologically meaningful features from sparse datasets. More recently, attention-based mechanisms have further improved interpretability by highlighting the most informative genes or pathways, making predictions more clinically transparent. Together, these architectural innovations mitigate key challenges in gene expression analysis, such as high dimensionality, sparsity, heterogeneity, and lack of interpretability, while enhancing robustness and classification performance. Neural network designs are being used to categorize Using gene expression data to determine cancer subtypes, enhancing diagnostic accuracy and personalizing treatment regimens. These architectures, such as convolutional and RNN, extract significant Information underlying protein expression

profiles enable the differentiation between cancer subtypes. Attention mechanisms and transfer learning procedures improve model interpretability by identifying key genes linked with distinct subtypes. Overall, this paradigm illustrates neural networks' potential to advance precision oncology by robustly classifying cancer Subtypes employing transcript expression information, as seen in the Figure 4 [46].

Liu et al. [47] conducted a comprehensive bioinformatics study using artificial neural networks (ANNs) to identify characteristic genes associated with cervical cancer (CC). By analyzing RNA sequencing data from multiple GEO datasets, the study identified differentially expressed genes (DEGs) between normal and cancerous cervical tissues. The authors applied random-forest filtering and established a neural network model using these characteristic genes, with Cox regression employed to verify the predictive accuracy. The proposed approach offers several benefits, including robust prediction of CC, insights into molecular mechanisms, identification of potential biomarkers, and guidance for immunotherapeutic interventions. However, limitations include the reliance on public datasets without experimental validation, incomplete understanding of viral and tumor immune escape mechanisms, and the need to consider epigenetic and immune regulatory factors, highlighting the necessity for further studies to validate and expand these findings.

Ren et al. [48] proposed a Multi-view Graph Neural Network (MVGNN) to classify breast cancer differentiation and subtypes by integrating multi-omics data (gene expression, DNA methylation, and CNV). The framework constructs weighted patient similarity networks for each omics type, applies Graph Convolutional Networks (GCN) to learn features, and employs an attention mechanism to fuse multi-omics representations. This model was designed to overcome the limitations of single-omics and traditional ML methods that often bias predictions toward one data type. Experimental validation on TCGA datasets demonstrated that MVGNN achieved superior performance in both binary and multi-class breast cancer classification compared to baseline models. The main benefits include robust multi-omics integration and improved accuracy, but its reliance on extensive preprocessing and the complexity of heterogeneous data fusion pose challenges for clinical adoption and scalability.

Zhou et al [49], The Nottingham Prognostics Index (NPI) is a prognostic metric designed for predicting mortality in treatable basic cancer of the breast. With advances in next-generation sequencing, multi-omics data collection allows for the examination of a wide range of physiological measurements to gain a better knowledge of disease progression. This work sought to find multi-omics indicators linked to breast cancer prognosis and survival, as well as to create a prediction model for several NPI classes. The suggested model performed exceptionally well, with an accuracy of 98.48% and an area under the curve (AUC) of 0.9999. The findings demonstrate substantial connections between the collected omics data and breast cancer prognosis and survival, highlighting

Table 3. Comparison of Literature Done in Neural Network Based on Approaches

Author/ Reference	Technique	Significance	Limitation
Liu et al. [47]	ANN with RF filtering	Identifies characteristic genes of cervical cancer, provides robust prediction, insights into molecular mechanisms, biomarker discovery, and guidance for immunotherapy.	Relies on public datasets without experimental validation; incomplete understanding of viral/tumor immune escape; epigenetic and immune regulatory factors not fully considered.
Ren et al. [48]	MVGNN	Robust integration of multi-omics data; improved breast cancer classification.	Requires extensive preprocessing; complex for clinical use.
Zhou et al [49]	High-dimensional embedding	Reduces dimensionality by embedding multi-omics features into a compact space, enabling the discovery of survival biomarkers and improving breast tumor prediction.	Dependence on multi-omics data limits applicability to less comprehensive datasets.
Yin et al [50]	Multi-omics graph convolutional network (M-GCN)	Captures complex biological interactions by modeling gene–gene relationships within graph structures, enhancing molecular subtyping accuracy through integrated multi-omics learning.	Computationally intensive due to integration of high-dimensional multi-omics data.
Choi et al [51]	moBRCA-net	Combines DL with self-attention to prioritize biologically significant features, improving prediction accuracy and supporting targeted treatment design.	Limited performance with incomplete datasets, while self-attention increases computational demands.
Allogmani et al. [52]	CPLDC-AOATL combining bilateral filtering, Inception-ResNetv2, and BiLSTM	Provides rapid, automated, and highly accurate detection of cervical precancerous lesions from medical images; aids early diagnosis and treatment with 99.53% accuracy.	Relies on specific pre-trained models and datasets; requires incorporation of multi-modal data and improved interpretability for clinical adoption.
Amin et al. [53]	Multimodal DL with CNNs	High accuracy in NSCLC classification using RNA-seq, miRNA-seq, and WSIs.	Limited to selected modalities; generalization across cancers is challenging.
Kesimoglu et al [54]	SUPREME	Employs a hybrid DL pipeline that enhances subtype prediction accuracy by integrating multiple omics signals into a unified model.	Relies on availability of comprehensive datasets.
Guo et al [55]	BCDForest DL model	Designed to work with small-scale biological datasets by combining ensemble methods with DL, reducing overfitting and improving generalizability.	May underperform when applied to datasets with very different structures.
Geeitha et al. [56]	Bi-RNN with HSIC and ensemble ML classifiers	Predicts cervical cancer recurrence and survival; enables early risk stratification, biomarker identification, and preventive interventions.	Dependent on retrospective datasets; requires real-time clinical validation; integration with additional genomic/clinical data needed for broader applicability.

biomarkers such as CDCA5, IL17RB, MUC2, NOD2, and NXPH4 in the gene expression dataset. Along with MED30, RAD21, EIF3H, and EIF3E from the copy number data. However, the reliance on multi-omics data may limit applicability to less comprehensive datasets, and high-dimensional analysis can increase complexity, risking overfitting.

Yin et al [50], proposed multi-omics graph convolutional network (M-GCN) is a new molecular subtyping system that uses robust graph convolutional networks to incorporate multi-omics data. To choose transcriptome features linked with molecular subtypes, the framework uses the Hilbert-Schmidt independence criterion of least absolute shrinkage and selection operator (HSIC Lasso). It then generates multi-view representations of samples using gene expression, single nucleotide variants (SNV), and copy number variation (CNV) data. The M-GCN model surpasses existing techniques for classifying breast and stomach cancers, and its identified subtype-specific biomarkers correspond to clinical knowledge, meaning accurate diagnosis and targeted treatment development. However, the reliance on graph structures may limit the model's ability to capture complex biological interactions, while integrating multi-omics data can increase computational complexity.

Choi et al [51], presented moBRCA-net, DL- Cancer

of the breast category identification framework utilizing multi-omics information. It takes into account biological interactions when combining data on expression of genes, DNA methylation, and microRNA expression. Each dataset is processed by a self-attention module, which determines the significance of each feature. The features are translated into new representations, which enable moBRCA-net to predict subtypes. Despite, The model's dependence on multi-omics data may limit applicability with incomplete datasets, and self-attention increases computational demands.

Allogmani et al. [52] proposed an enhanced method for detecting and classifying cervical precancerous lesions using the Archimedes Optimization Algorithm with Transfer Learning (CPLDC-AOATL). The approach integrates bilateral filtering for image denoising, Inception-ResNetv2 for feature extraction, AOA for hyperparameter tuning, and a BiLSTM model for classification. Tested on benchmark medical image datasets, the CPLDC-AOATL method achieved a high accuracy of 99.53%, outperforming existing techniques. The benefits include rapid, automated, and highly accurate detection of cervical cancer from images, aiding early diagnosis and treatment. Limitations include reliance on specific pre-trained models and datasets, with future work needed to incorporate multi-modal data and improve interpretability for clinical use.

Table 4. Comparison of Literature Done in Transfer Learning based on Approaches

Author/Reference	Technique	Significance	Limitation
Tabassum et al. [58]	Dimensionality reduction and XAI with ensemble classifiers (Logistic Regression, SVM, XGBoost) on mRNA gene expression data	· Accurately classifies 33 cancer types, identifies biologically meaningful biomarkers. · Reduces computational cost, and enables personalized treatment.	Dependent on pre-processed gene expression datasets; requires clinical validation for real-world applicability.
Franchini et al [59]	Single-cell gene set enrichment analysis (scGSEA) and single-cell mapper (scMAP)	· Enhances single-cell analysis by accurately identifying gene activity patterns · Aids in mapping novel cell profiles, improving cancer research insights.	Require extensive computational resources for high-dimensional single-cell data analysis.
Ming et al [60]	DCE-MRI	· Enhances predictive accuracy for HR status and PAM50 molecular. · Provides a gene expression-based approach	Potentially affecting comprehensive subtype classification.
Pan et al. [61]	Trans-PtLR (Robust Transfer Learning with t-distributed errors)	· Improves prediction accuracy in gene expression by handling outliers and heavy-tailed data.	Computationally intensive; performance depends on source dataset quality.
Samee et al [62]	(GN-AlexNet)	· Demonstrates significant improvements in accuracy · Enhance feature extraction and classification performance.	Limit its deployment in resource-constrained environments or for real-time applications.
Muhammad et al [63]	DRNet	· Enhances accessibility for healthcare professionals · Supporting early diagnosis and intervention.	Require retraining on diverse datasets to maintain diagnostic accuracy across varied breast cancer subtypes.
Zhang et al [64]	T-GEM	· providing insights into biological functions and potential biomarkers · Supporting targeted cancer research and diagnostics.	T-GEM have limitations in handling large-scale gene expression data with complex interactions, potentially impacting its scalability for diverse cancer types.
Wang et al [65]	Graph-based deep embedding clustering (GDEC)	· Highlighting its enhanced analytical capability · Improvements in clustering accuracy	With high-dimensional noise in diverse scRNA-seq datasets.
Attallah et al [66]	CerCan·Net	· CerCan·Net achieves high accuracy for cervical cancer diagnosis	Face limitations in generalizability across diverse cervical cancer subtypes and varied dataset conditions.
Srikantamurthy et al [67]	hybrid CNN-LSTM model	· Cancer subtype differentiation in diagnostic applications. · Achieves high accuracy in breast cancer histopathology classification,	Limited generalizability across diverse histopathology datasets due to variability in imaging conditions

Amin et al. [53] introduced a multimodal DL framework for non-small cell lung cancer (NSCLC) classification, integrating RNA-seq, miRNA-seq, and whole-slide images (WSIs). Unlike earlier approaches that either used traditional ML on molecular data or single-modality deep learning, this study leveraged CNNs across multiple modalities. Experimental results demonstrated high classification performance, with accuracies of 96.79% (RNA-seq), 98.59% (miRNA-seq), and 89.73% (WSIs), alongside strong F1-scores and AUC values, surpassing prior state-of-the-art results. The study highlights CNNs' capability to handle high-dimensional omics data and large-scale pathology images, enabling early-stage cancer detection and precise subtype classification. While results are promising, challenges remain in extending the approach to additional modalities (e.g., DNA methylation, CNV) and generalizing across different cancer types.

Kesimoglu et al [54], proposed SUPREME is a subtype prediction technology created by analysing multiomics data and patient relationships using network convolutions on different patient matching matrices. It creates patient embedding based on all multiomics features and incorporates all potential combinations

to capture complimentary signals. The strategy beat previous integrated cancer prediction tools and baseline methodologies across three datasets from The Cancer Genome Atlas (TCGA) and the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC). SUPREME-inferred subtypes had considerably greater survival rates compared to nine cancer subtype differentiating tools and baseline techniques. The findings show that, with the correct combination of datatypes and patient relationships, SUPREME can identify hidden properties in tumor subgroups that generate substantial mortality variations and enhance the fundamental reality identify that is mostly based on an individual datatype. However, the model's reliance on multiomics data may restrict its utility in situations when such extensive datasets are absent.

Basaad et al. [55] suggested GraphX-Net, a Shapley Value-based Graph Neural Network (GNN) framework for predicting breast cancer relapse. The model applies graph convolutional layers to learn node embeddings and uniquely integrates Shapley values to quantify feature contributions, manage node thresholds, and capture neighboring effects, thereby improving interpretability

alongside prediction accuracy. Through this combination, GraphX-Net forms distinct patient clusters and offers transparent insights into risk factors influencing relapse, bridging the gap between accuracy and explainability. Experimental evaluation on the METABRIC cohort confirmed state-of-the-art performance, while also enabling visualization of graph connectivity and feature importance.

Geeitha et al. [56] developed a bidirectional recurrent neural network (Bi-RNN) model to predict cervical cancer recurrence and patient survival by integrating clinical risk factors and lncRNA gene signatures. Clinical features were analyzed using Random Forest, Logistic Regression, Gradient Boosting, and SVM, with Random Forest achieving 91.2% precision. The Hilbert-Schmidt Independence Criterion (HSIC) linked lncRNA signatures with protein-coding genes to identify biomarkers associated with recurrence. The Bi-RNN model effectively predicted recurrence and survival, enabling early risk stratification and targeted interventions. Benefits include accurate prognosis, biomarker discovery, and preventive guidance, while limitations involve reliance on retrospective datasets and the need for real-time clinical validation (Table 3).

Recent research demonstrates that neural network architectures provide complementary strategies to address the intrinsic challenges of gene expression data in cancer diagnostics. Convolutional models, such as bio-inspired CNNs, effectively capture local co-expression patterns while reducing noise. Self-explainable frameworks like Deep GONet enhance interpretability by linking gene features to phenotypes. Embedding and graph-based approaches, including high-dimensional embeddings and M-GCN, reduce dimensionality and model complex gene-gene interactions across multi-omics inputs. Optimization-driven RNNs capture sequential dependencies in regulatory pathways, while attention-based networks such as moBRCA-net highlight biologically significant features, bridging predictive performance with clinical interpretability. Ensemble and hybrid methods (e.g., BCDForest, SUPREME) improve robustness, particularly for small or heterogeneous datasets. Taken together, these advances underscore how neural networks are not only improving accuracy but also tackling key issues of dimensionality, heterogeneity, and clinical usability, positioning them as critical tools for next-generation cancer subtype classification.

3.4 Transfer learning based on approaches

A transfer of ownership method of learning for tumor subtype categorization using expression of genes information uses pre-trained models to improve subtype identification accuracy and efficiency [57]. Using knowledge obtained from large datasets, this strategy successfully adapts existing models to new, perhaps smaller or more specific gene expression datasets, allowing for enhanced classification performance with less training data. The system includes approaches such as fine-tuning and feature extraction, allowing the model to capture significant biological signals while overcoming

the obstacles of overfitting and data sparsity. This strategy not only speeds up the model training process, but it also improves generalizability across many cancer types, thereby helping precision medicine initiatives by allowing for accurate and quick cancer subtype classification as shown in Figure 5.

Tabassum et al. [58] proposed a precision cancer classification framework using mRNA gene expression data, combining dimensionality reduction, feature selection, and Explainable AI (XAI) techniques. The pipeline reduces the original 19,238-gene dataset to 500 key features while retaining critical information, and employs an ensemble of Logistic Regression, SVM, and XGBoost classifiers to achieve 96.61% accuracy across 33 cancer types. Explainable AI, via SHAP scores, identifies and prioritizes cancer-specific biomarker genes, validated against Differential Gene Expression (DGE) analysis. The benefits include accurate, rapid cancer classification, reduced computational cost, and identification of biologically meaningful biomarkers for personalized treatment. Limitations involve reliance on preprocessed gene expression datasets and the need for clinical validation to confirm biomarker utility in real-world scenarios.

Franchini et al [59], Researchers have created two novel approaches for studying single-cell gene set enrichment analysis (scGSEA) and single-cell mapper (scMAP). ScGSEA detects coordinated gene activity at the single-cell level using latent data representations and gene set enrichments, whereas scMAP uses transfer learning techniques to repurpose and situate freshly created cells into the standard cell the atlas database. Both approaches can accurately duplicate repeating patterns of pathway activation exhibited by cells under multiple experimental conditions, as well as place and contextual novel a single-cell characteristics on the breast tumor map. May, require extensive computational resources for high-dimensional single-cell data analysis.

Ming et al [60], researched DCE-MRI technology Using tumor and peri-tumor region slice images, AI models can predict HR status and PAM50 molecular

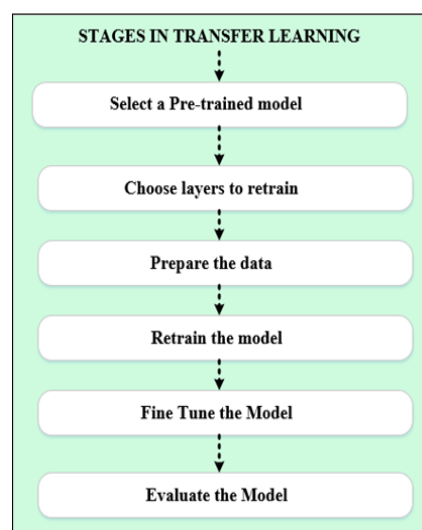


Figure 5. Stages in Transfer Learning

subtypes of breast cancer. Inception-v3 with Exception systems performed better overall, although there was little variation in ER status or PAM50 classes. Comparisons with previous studies revealed that the majority of current investigations predict IHC-based molecular subtypes, but our algorithms predicted gene expression-based chemical types, notably PAM50 fundamental types, providing a better knowledge of illness characteristics. The models were compared to previous radionics papers. However, Limited in differentiating ER status and specific PAM50 subtypes, potentially affecting comprehensive subtype classification.

Pan et al. [61] introduced a robust transfer learning framework (Trans-PtLR) for integrating multi-source gene expression data under high-dimensional linear regression. Unlike traditional approaches that assume normal error distribution, Trans-PtLR incorporates t-distributed errors, enabling robustness against outliers and heavy-tailed data, which are common in genomic datasets. Their three-step algorithm combines penalized maximum likelihood estimation with transferable source selection, effectively preventing negative transfer. Applied to GTEx datasets, the model demonstrated superior accuracy in predicting gene expression (e.g., JAM2 gene across multiple brain tissues) compared to conventional transfer learning methods. This advancement highlights the potential of robust transfer learning in capturing complex gene regulatory patterns, thereby supporting improved prediction in gene-expression-based cancer subtype analysis.

Samee et al [62], investigated a hybrid deep transfer learning (GN-AlexNet) model for BT tri-classification (pituitary, meningioma, and glioma). The proposed solution combines GoogleNet architecture with the AlexNet model, removing GoogleNet's five layers and adding ten layers from the AlexNet model to automatically extract and classify attributes. On the same CE-MRI dataset, the proposed model was compared to transfer learning techniques (VGG-16, AlexNet, SqueezNet, ResNet, and MobileNet-V2) and ML/DL. The proposed model outperformed current strategies in terms of both accuracy and sensitivity (99.51 percent and 98.90 percent, respectively). However, the model's complexity may limit its deployment in resource-constrained environments or for real-time applications.

Muhammad et al [63], described a new deep feature extraction DRNet model for detecting Identification of breast cancer subtypes using the Breakhis data. The model uses a transfer learning technique, which relies on trained multilayer artificial neural networks that have little storage and processing capability. The model outperforms certain previous literature studies and is capable of extracting deep features from raw histology pictures of breast cancer. This methodology allows medical professionals to acquire quick and precise findings, identifying treatments for various Breast tumor types are identified early on, preserving lives and expenditures. May, require retraining on diverse datasets to maintain diagnostic accuracy across varied breast cancer subtypes.

Zhang et al [64], researched a new DL architecture

known as T-GEM, or Transformer for Gene Expression Modelling, which is beneficial for predicting cancer-related phenotypes. The model employs a thorough learning method, with the first layer focused on gene-gene interactions and subsequent layers focusing on phenotype-related genes. T-GEM's self-attention can identify biological functions linked to expected phenotypes. The researchers also devised a method for extracting the regulatory network, which identified network hub genes as potential markers for expected symptoms. However, T-GEM have limitations in handling large-scale gene expression data with complex interactions, potentially impacting its scalability for diverse cancer types.

Wang et al [65], researched introduces graph-based deep embedding clustering (GDEC), a method for grouping scRNA-seq data using Transferring knowledge between animals and batch. GDEC utilizes convolutional networks based on graphs to overcome sparse gene expression matrices and split groups of cells into a space with fewer dimensions, therefore decreasing noise effects. The method builds a model from existing scRNA-seq datasets and fine-tunes it with transfer learning techniques. GDEC was used to uterine fibroids scRNA-seq data, exposing a new cell type and uncovering new routes between different cell types, demonstrating its improved analytical capabilities. The researchers also performed cross-species and cross-batch clustering investigations. However, Graph-based deep embedding clustering enhances analytical capability and clustering accuracy but may face limitations with high-dimensional noise in diverse scRNA-seq datasets.

Attallah et al [66], introduced CerCan·Net, an effective computer-assisted diagnosis tool for cervical cancer. It uses three lightweight CNNs that have fewer parameters and deeper layers than prior models: Mobile Net, DarkNet-19, and ResNet-18. CerCan·Net utilizes transfer learning to extract deep features from the final three levels of each CNN, including input from many layers. It then investigates the impact of developing a smaller set of deep features to distinguish various subgroups of cervical cancer. CerCan·Net achieves 97.7% and 100% accuracy for SIPaKMeD and Mendeley datasets, respectively, with 400 and 200 features. Its superior performance in comparison to contemporary CADs makes it appropriate for cytopathologists in automated inspection, avoiding limitations in ordinary diagnostic. Despite, CerCan·Net, while highly accurate, face limitations in generalizability across diverse cervical cancer subtypes and varied dataset conditions.

Srikantamurthy et al [67], proposed hybrid CNN-LSTM model was tested against existing models for breast histopathology image categorization, such as VGG-16, ResNet50, and Inception. The Adam optimizer was determined to be the most accurate and cause the least amount of model loss. The model has the best overall accuracy of 99% for binary classification of benign and malignant cancer, and 92.5% for classification into multiple classes of benign and dangerous cancer subcategories, respectively. However, limited generalizability across diverse histopathology datasets due to variability in

imaging conditions (Table 4).

Advanced models in cancer research have improved diagnostic accuracy and feature extraction. DEGnext enhances prediction accuracy by identifying important gene expressions, whereas single-cell analysis methods such as scGSEA and scMAP precisely detect gene activity patterns and map cell profiles. DCE-MRI and T-GEM improve molecular categorization and gene function insights, although they have scalability limitations. The ATRCN model improves liver cancer prognosis by identifying prognostic markers, although its applicability may be limited across different cancer types. GN-AlexNet and DRNet provide better accuracy and accessibility for early diagnosis, however deployment in resource-constrained situations or with different cancer subtypes may necessitate changes. GDEC's deep embedding clustering and CerCan-Net are highly accurate in scRNA-seq and cervical cancer diagnosis, but may be limited by dataset diversity.

3.5 Critical Synthesis of Gaps

The comparative analysis of ML, explainable AI, neural network, and transfer learning approaches highlights several recurring limitations that continue to hinder the clinical applicability of cancer subtype classification models. These gaps can be broadly grouped into three categories: data-related, model-related, and clinical translation.

- **Data-related gaps:** Gene expression and multi-omics datasets remain fragmented, heterogeneous, and often limited in sample size. The lack of standardized protocols for data collection, preprocessing, and normalization leads to inconsistencies across studies, making reproducibility and cross-comparison challenging. Small and imbalanced datasets also increase the risk of biased models that fail to generalize across populations.

- **Model-related gaps:** While DL and neural networks capture complex, non-linear biological relationships, they are computationally intensive, prone to overfitting, and difficult to scale for large, high-dimensional omics datasets. ML models such as SVMs and decision trees often provide interpretability but struggle with the complexity of modern multi-omics data. Even when models achieve high accuracy in controlled settings, their robustness and scalability remain limited when applied to diverse real-world datasets.

- **Clinical translation gaps:** Despite progress in explainable AI, many models still operate as "black boxes," limiting their clinical interpretability and trustworthiness. Moreover, the absence of cross-institutional validation and adaptation mechanisms restricts their deployment across varied healthcare systems. The lack of integration with clinical workflows further slows the transition from research prototypes to practical diagnostic tools.

Taken together, these cross-cutting gaps indicate that no single methodological category is sufficient on its own. A unified framework that balances predictive accuracy, computational efficiency, and interpretability is required to ensure clinically viable solutions. Future directions should prioritize integrating multi-omics data, advancing

explainable AI techniques, optimizing algorithms for low-resource environments, and employing transfer learning for robust generalization across populations. This synthesis sets the stage for the conceptual framework proposed in the following section, which envisions a pathway toward scalable, interpretable, and patient-centered cancer subtype classification.

4. Summary and Discussion

This review highlights the limitations and Future Prospects in Tumor Subtype Identification Utilizing Expression Gene Databases, underscoring the potential of ML and explainable AI frameworks to transform cancer diagnostics. However, challenges remain in data quality, interpretability, computational demands, generalizability, tumor heterogeneity, and standardization. Addressing these issues requires integrating multi-omics data, advancing XAI techniques, optimizing computational methods, leveraging transfer learning, and creating standardized protocols. Future research can improve the robustness, scalability, and clinical utility of ML models, contributing to more accurate cancer subtype classification and improved patient outcomes.

- The lack of standardized practices in data collection, pre-processing, and model evaluation in cancer research studies hinders comparability across studies.

- Advanced pre-processing and data normalization techniques can improve data reliability by addressing variability in gene expression data quality and inconsistencies across diverse sources

- Advance XAI techniques aim to enhance the accessibility, interpretability, and clinical usefulness of complex ML models, particularly deep learning, in clinical practice.

- Optimizing algorithms can improve computational efficiency and scalability of DL models, especially in low-resource settings.

- To enhance model robustness across different cancers and populations, it is recommended to expand the use of transfer learning and diversify training datasets.

- Integrating multi-omics data and developing personalized models can enhance subtype accuracy by better accounting for individual tumor characteristics, addressing the challenges faced by many models.

- **Dataset biases and representational limitations:** Many widely used datasets, such as The Cancer Genome Atlas (TCGA), are disproportionately composed of samples from specific geographic regions, ethnic groups, and clinical settings. This demographic skew introduces biases that can limit the generalizability of models to underrepresented populations. Moreover, variations in sequencing protocols, institutional standards, and data curation practices can compound these disparities. Future research should address these limitations by incorporating more diverse, multi-institutional datasets and developing fairness-aware learning approaches that explicitly account for population heterogeneity.

- **Clinical integration of AI-driven cancer subtype classifiers** faces significant hurdles. Regulatory approval processes (e.g., FDA, EMA) require rigorous validation,

standardized reporting, and clear evidence of safety and efficacy. Equally important is physician acceptance: models must provide interpretable outputs that clinicians can trust, while also integrating smoothly into existing diagnostic workflows and electronic health records (EHRs). Without addressing these barriers, even technically strong models may fail to achieve real-world impact.

By addressing these limitations with targeted advancements, Future research can increase the practical impact and clinical usefulness of disease subgroup categorization using ML approaches by addressing these constraints through targeted breakthroughs.

To address these cross-cutting gaps, we propose an integrative framework for AI-driven cancer subtype classification. This framework positions ML as the foundation for baseline classification, neural networks as advanced engines for high-dimensional data, XAI as the interpretability layer to bridge clinical usability, and transfer learning as the adaptability mechanism to ensure generalization across datasets and populations. Figure 6 illustrates this unified perspective.

4.1 Translational Challenges and Solutions

While technical advancements in ML, NN, XAI, and TL have significantly improved cancer subtype classification, the translation of these methods into clinical practice faces major challenges:

- **Regulatory Hurdles** – AI-based diagnostic systems must meet stringent requirements from regulatory bodies such as the FDA and EMA. Demonstrating robustness, reproducibility, and patient safety across diverse populations is essential before clinical approval.
- **Physician Acceptance and Trust** – Clinicians require models that are not only accurate but also interpretable. Black-box systems, even with high predictive power, often face resistance due to the lack of transparent decision-making processes. XAI-driven approaches offer a pathway to bridging this gap.
- **Workflow Integration** – AI tools must be seamlessly integrated into existing clinical infrastructures, such as electronic health records (EHRs) and hospital IT systems, without creating additional burden on healthcare providers.
- **Ethical and Legal Considerations** – Liability in cases of misdiagnosis, protection of patient privacy, and compliance with data-sharing regulations are critical issues that must be addressed before deployment.

Proposed Solutions

- Develop standardized reporting protocols to improve reproducibility and regulatory acceptance.
- Co-design AI tools with physicians to enhance trust and usability.
- Employ federated learning and secure multi-institutional data-sharing frameworks to overcome data heterogeneity while preserving privacy.
- Incorporate explainability modules (XAI) and uncertainty quantification in models to improve physician confidence in clinical decision-making.

By explicitly addressing these translational barriers,

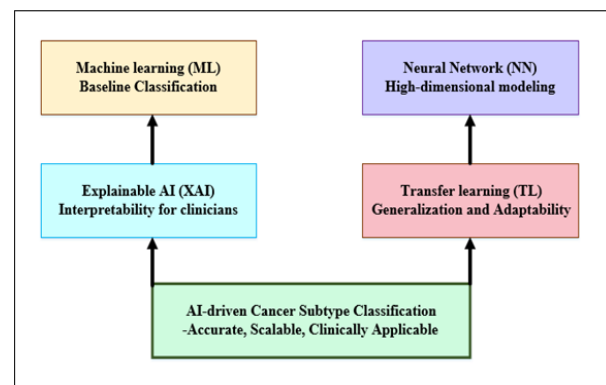


Figure 6. Conceptual Framework for AI-driven Cancer Subtype Classification

future research can ensure that AI-driven cancer subtype classification not only achieves technical excellence but also delivers clinically deployable, ethically sound, and patient-centered solutions.

In conclusion, the article highlights the potential for change of neural networks. And explainable AI techniques in cancer subtype classification using gene expression data. While current advancements offer improved diagnostic accuracy, interpretability, and feature extraction, significant challenges remain, including issues with data quality, computational scalability, model generalizability, and the complexity of tumor heterogeneity. To overcome these limitations, future studies should prioritize the incorporation of multi-omics data, and the development of sophisticated preliminary processing and normalization tools, optimization of algorithms for low-resource settings, and expansion of transfer learning frameworks to enhance model robustness across diverse cancer types and populations. Additionally, creating standardized protocols for data collection, pre-processing, and evaluation will be essential to improve comparability and reproducibility across studies. By focusing on these advancements, the field can work toward AI-driven solutions that are more accurate, scalable, and clinically applicable, ultimately Contributing to better outcomes for patients with a more tailored cancer therapy strategy. However, for AI frameworks to transition from research to clinical deployment, regulatory compliance, physician acceptance, and seamless integration into healthcare systems remain critical. Future studies should incorporate validation strategies aligned with medical regulations, engage clinicians in co-designing interpretable systems, and develop deployment pathways compatible with hospital IT infrastructures. Only then can these AI solutions move beyond academic promise to deliver tangible benefits in clinical oncology. The distinctive contribution of this review lies in its critical synthesis of recurring challenges and the introduction of a conceptual framework uniting ML, NN, XAI, and TL. This framework offers a roadmap for future research to move beyond isolated methods toward integrated, scalable, and clinically relevant cancer diagnostic systems.

Declaration

Conflict of interest

On behalf of all authors, the corresponding author stated that they have no conflict of interest

Author Contribution

All authors are equally contributed

Ethical Approval

No humans or Animals were included in this work.

Consent for publication

All authors are given their consent for publication

References

1. Rawat V, Singh DP, Singh N, Kumar S. Detection and prognosis of cancer using state-of-the-art technologies: review, issues, and motivation. In 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 679-686). IEEE. 2023.
2. Abuya T. Lung Cancer Prediction from Elvira Biomedical Dataset Using Ensemble Classifier with Principal Component Analysis. *Journal of Data Analysis and Information Processing*. 2023 01 01;11:175-199. <https://doi.org/10.4236/jdaip.2023.112010>
3. Perelo LW, Gabernet G, Straub D, Nahnsen S. How tool combinations in different pipeline versions affect the outcome in RNA-seq analysis. *NAR genomics and bioinformatics*. 2024 03;6(1):lqae020. <https://doi.org/10.1093/nargab/lqae020>
4. Long Z, Li Y, Gan Y, Zhao D, Wang G, Xie N, Lovnicki JM, et al. ERRATUM: Roles of the HOXA10 gene during castrate-resistant prostate cancer progression. *Endocrine-Related Cancer*. 2025 06 01;32(6):x180465. <https://doi.org/10.1530/ERC-18-0465e>
5. Turki F, Khadem A, Jalali Aghchai A. Classification of Five Types of Cancer Data Based on Neural Network Methods, and Analysis of Gene Expression using the Feature Selection Fusion Method. *Iranian Journal of Mechanical Engineering Transactions of ISME*. 2024 08 22;26(2):46-66. <https://doi.org/10.30506/ijmep.2023.2003519.1940>
6. Cheng C, Liang Z, Xu R, Gu Y, Wu SW, Wang H, Shi N, Zhang Q, Tao Y, Li W. Luadnet: A Deep Learning Model For Prediction Of Clinical Outcomes in Lung Adenocarcinoma Based on Gene Expression Signatures. 2025 06 17;. <https://doi.org/10.2139/ssrn.5290312>
7. Soeparmi S, Yunianto M, Amalia LR. Lung Cancer Classification using Gray-Level Co-Occurrence Matrix Feature Extraction and Forward Selection Feature Selection based on the K-Nearest Neighbor Algorithm. *Indonesian Journal Of Applied Physics*. 2025 05 04;15(1):133-146. <https://doi.org/10.13057/ijap.v15i1.90378>
8. Hassan A, Naeem S, Eldosoky M, Mabrouk M. MACHINE Learning Approaches For Integrative Multi-Omics Data Analysis Of Cancer: A Systematic Review. *Biomedical Engineering: Applications, Basis and Communications*. 2025 06 17;. <https://doi.org/10.4015/S1016237225300044>
9. Nkurunziza F, Kabanda R, McSharry P. Enhancing poverty classification in developing countries through machine learning: a case study of household consumption prediction in Rwanda. *Cogent Economics & Finance*. 2025 Dec 31;13(1):2444374. <https://doi.org/10.1080/23322039.2024.2444374>
10. Aziz MMH, Mahmood SA. Utilizing Machine Learning Techniques for Cancer Prediction and Classification based on Gene Expression Data. *UHD Journal of Science and Technology*. 2025 06 02;9(1):135-148. <https://doi.org/10.21928/uhdst.v9n1y2025.pp135-148>
11. Liu W, He H, Chicco D. Gene signatures for cancer research: A 25-year retrospective and future avenues. *PLoS computational biology*. 2024 Oct;20(10):e1012512. <https://doi.org/10.1371/journal.pcbi.1012512>
12. Huang D, Li Z, Jiang T, Yang C, Li N. Artificial intelligence in lung cancer: current applications, future perspectives, and challenges. *Frontiers in Oncology*. 2024;14:1486310. <https://doi.org/10.3389/fonc.2024.1486310>
13. Scientific LL. Unraveling Lung Cancer Through Genomic Insights And Ensemble Deep Learning. *Journal of Theoretical and Applied Information Technology*. 2024;102(2).
14. Harishbhai Tilala M, Kumar Chenchala P, Choppadandi A, Kaur J, Naguri S, Saoji R, Devaguptapu B. Ethical Considerations in the Use of Artificial Intelligence and Machine Learning in Health Care: A Comprehensive Review. *Cureus*. 2024 06;16(6):e62443. <https://doi.org/10.7759/cureus.62443>
15. Amanzholova A, Coşkun A. Enhancing cancer stage prediction through hybrid deep neural networks: a comparative study. *Frontiers in Big Data*. 2024;7:1359703. <https://doi.org/10.3389/fdata.2024.1359703>
16. Das A, Chatterjee S. Cancer Classification Based on an Integrated Clustering and Classification Model Using Gene Expression Data. 2022 01 01:461-470.
17. Nethala TR, Sahoo BK, Srinivasulu P. Optimal gene therapy network: Enhancing cancer classification through advanced AI-driven gene expression analysis. *e-Prime - Advances in Electrical Engineering, Electronics and Energy*. 2024 03 01;7:100449. <https://doi.org/10.1016/j.prime.2024.100449>
18. Abidalkareem A, Ibrahim AK, Abd M, Rehman O, Zhuang H. Identification of Gene Expression in Different Stages of Breast Cancer with Machine Learning. *Cancers*. 2024 05 14;16(10):1864. <https://doi.org/10.3390/cancers16101864>
19. Babichev S, Liakh I, Kalinina I. Applying the Deep Learning Techniques to Solve Classification Tasks Using Gene Expression Data. *IEEE Access*. 12:28437-28448. <https://doi.org/10.1109/ACCESS.2024.3368070>
20. Al-Azani S, Alkhnbashi OS, Ramadan E, Alfarraj M. Gene Expression-Based Cancer Classification for Handling the Class Imbalance Problem and Curse of Dimensionality. *International Journal of Molecular Sciences*. 2024 02 09;25(4):2102. <https://doi.org/10.3390/ijms25042102>
21. Khalsan M, Mu M, Al-Shamery ES, Machado L, Ajit S, Agyeman MO. Fuzzy Gene Selection and Cancer Classification Based on Deep Learning Model. *arXiv preprint arXiv*. 2023 05 04;. <https://doi.org/10.48550/arXiv.2305.04883>
22. Joshi P, Basso B, Wang H, Hong S, Giardina C, Shin D. rPAC: Route based pathway analysis for cohorts of gene expression data sets. *Methods (San Diego, Calif.)*. 2022 02;198:76-87. <https://doi.org/10.1016/j.ymeth.2021.10.002>
23. Nahiduzzaman M, Abdulrazak LF, Ayari MA, Khandakar A, Islam SMR. A novel framework for lung cancer classification using lightweight convolutional neural networks and ridge extreme learning machine model with SHapley Additive exPlanations (SHAP). *Expert Systems with Applications*. 2024 08 15;248:123392. <https://doi.org/10.1016/j.eswa.2024.123392>
24. Simhadri N, Swamy TNVR. Awareness among teaching on AI and ML applications based on fuzzy in education sector at USA. *Soft Computing*. 2023 05 12;. <https://doi.org/10.1007/>

- s00500-023-08329-z
25. Rukhsar L, Bangyal WH, Ali Khan MS, Ag Ibrahim AA, Nisar K, Rawat DB. Analyzing RNA-Seq Gene Expression Data Using Deep Learning Approaches for Cancer Classification. *Applied Sciences*. 2022 01;12(4):1850. <https://doi.org/10.3390/app12041850>
 26. Sun P, Wu Y, Yin C, Jiang H, Xu Y, Sun H. Molecular Subtyping of Cancer Based on Distinguishing Co-Expression Modules and Machine Learning. *Frontiers in Genetics*. 2022;13:866005. <https://doi.org/10.3389/fgene.2022.866005>
 27. Park J, Rhee J. Integrative Analysis of ATAC-Seq and RNA-Seq through Machine Learning Identifies 10 Signature Genes for Breast Cancer Intrinsic Subtypes. *Biology*. 2024 Oct 07;13(10):799. <https://doi.org/10.3390/biology13100799>
 28. Babichev S, Yarema O, Savchenko A. Evaluating proximity metrics for gene expression data: A hybrid model integrating data mining and machine learning techniques for disease diagnosis systems. *Biomedical Signal Processing and Control*. 2025 Dec 01;110:108115. <https://doi.org/10.1016/j.bspc.2025.108115>
 29. Alanazi SA, Alshammari N, Alruwaili M, Junaid K, Abid MR, Ahmad F. Integrative analysis of RNA expression data unveils distinct cancer types through machine learning techniques. *Saudi Journal of Biological Sciences*. 2024 03;31(3):103918. <https://doi.org/10.1016/j.sjbs.2023.103918>
 30. Babichev S, Yasinska-Damri L, Liakh I. A Hybrid Model of Cancer Diseases Diagnosis Based on Gene Expression Data with Joint Use of Data Mining Methods and Machine Learning Techniques. *Applied Sciences*. 2023 01;13(10):6022. <https://doi.org/10.3390/app13106022>
 31. Liu Z, Liu L, Weng S, Guo C, Dang Q, Xu H, Wang L, et al. Machine learning-based integration develops an immune-derived lncRNA signature for improving outcomes in colorectal cancer. *Nature Communications*. 2022 02 10;13(1):816. <https://doi.org/10.1038/s41467-022-28421-6>
 32. Mohamed TIA, Ezugwu AE, Fonou-Dombeu JV, Ikotun AM, Mohammed M. A bio-inspired convolution neural network architecture for automatic breast cancer detection and classification using RNA-Seq gene expression data. *Scientific Reports*. 2023 09 05;13(1):14644. <https://doi.org/10.1038/s41598-023-41731-z>
 33. Sarkar S, Mali K. Breast Cancer Subtypes Classification with Hybrid Machine Learning Model. *Methods of Information in Medicine*. 2022 09;61(3-04):68-83. <https://doi.org/10.1055/s-0042-1751043>
 34. Senbagamalar L, Logeswari S. Genetic Clustering Algorithm-Based Feature Selection and Divergent Random Forest for Multiclass Cancer Classification Using Gene Expression Data. *International Journal of Computational Intelligence Systems*. 2024;17(1). <https://doi.org/10.1007/s44196-024-00416-9>
 35. Bharati S, Mondal MRH, Podder P, Kose U. Explainable Artificial Intelligence (XAI) with IoT for Smart Healthcare: A Review. *Interpretable Cognitive Internet of Things for Healthcare*. 2023;1-24.
 36. Wani NA, Kumar R, Bedi J. DeepXplainer: An interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence. *Computer Methods and Programs in Biomedicine*. 2024 01;243:107879. <https://doi.org/10.1016/j.cmpb.2023.107879>
 37. Li H, Han Z, Sun Y, Wang F, Hu P, Gao Y, Bai X, et al. CGMega: explainable graph neural network framework with attention mechanisms for cancer gene module dissection. *Nature Communications*. 2024 07 17;15(1):5997. <https://doi.org/10.1038/s41467-024-50426-6>
 38. Abhang MVK, Gunjal BL. Deep Graph Ensemble Convolutional Neural Networks for Drug Response Prediction from Multi- Omics Cancer Cell Lines. *COMPUTER*. 25(4).
 39. Sekaran K, Varghese RP, Krishnan S, Zayed H, El Allali A, Doss GPC. Dissecting Crucial Gene Markers Involved in HPV-Associated Oropharyngeal Squamous Cell Carcinoma from RNA-Sequencing Data through Explainable Artificial Intelligence. *Frontiers in Bioscience (Landmark Edition)*. 2024 06 18;29(6):220. <https://doi.org/10.31083/j.fbl2906220>
 40. Morabito F, Adornetto C, Monti P, Amaro A, Reggiani F, Colombo M, Rodriguez-Aldana Y, et al. Genes selection using deep learning and explainable artificial intelligence for chronic lymphocytic leukemia predicting the need and time to therapy. *Frontiers in Oncology*. 2023;13:1198992. <https://doi.org/10.3389/fonc.2023.1198992>
 41. Yang P, He C, Zhang P, Qin X, Zhang Q, Li D. MHGCN: A Multi-Channel Hybrid Graph Convolutional Neural Network for Cancer Drug Response Prediction. *IEEE transactions on computational biology and bioinformatics*. 2025;22(5):2241-2251. <https://doi.org/10.1109/TCBBIO.2025.3594056>
 42. Gutierrez-Chakraborty E, Chakraborty D, Das D, Bai Y. Discovering novel prognostic biomarkers of hepatocellular carcinoma using eXplainable Artificial Intelligence. *Expert Systems with Applications*. 2024 Oct 15;252(Pt B):124239. <https://doi.org/10.1016/j.eswa.2024.124239>
 43. Abuzinadah N, Kumar Posa S, Alarfaj AA, Alabdulqader EA, Umer M, Kim T, Alsubai S, Ashraf I. Improved Prediction of Ovarian Cancer Using Ensemble Classifier and Shaply Explainable AI. *Cancers*. 2023 Dec 11;15(24):5793. <https://doi.org/10.3390/cancers15245793>
 44. Altini N, Puro E, Taccogna MG, Marino F, De Summa S, Saponaro C, Mattioli E, Zito FA, Bevilacqua V. Tumor Cellularity Assessment of Breast Histopathological Slides via Instance Segmentation and Pathomic Features Explainability. *Bioengineering (Basel, Switzerland)*. 2023 03 23;10(4):396. <https://doi.org/10.3390/bioengineering10040396>
 45. Rajpal S, Rajpal A, Agarwal M, Kumar V, Abraham A, Khanna D, Kumar N. XAI-CNVMarker: Explainable AI-based copy number variant biomarker discovery for breast cancer subtypes. *Biomedical Signal Processing and Control*. 84:104979.
 46. Mall PK, Singh PK, Srivastav S, Narayan V, Paprzycki M, Jaworska T, Ganzha M. A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities. *Healthcare Analytics*. 2023 Dec 01;4:100216. <https://doi.org/10.1016/j.health.2023.100216>
 47. Liu L, Huang L, Deng L, Li F, Vannucci J, Tang S, Wang Y. Bioinformatics identification of characteristic genes of cervical cancer via an artificial neural network. *Chinese Clinical Oncology*. 2024 02;13(1):4. <https://doi.org/10.21037/cco-23-139>
 48. Ren Y, Gao Y, Du W, Qiao W, Li W, Yang Q, Liang Y, Li G. Classifying breast cancer using multi-view graph neural network based on multi-omics data. *Frontiers in Genetics*. 2024;15:1363896. <https://doi.org/10.3389/fgene.2024.1363896>
 49. Zhou L, Rueda M, Alkhateeb A. Classification of Breast Cancer Nottingham Prognostic Index Using High-Dimensional Embedding and Residual Neural Network. *Cancers*. 2022 02 13;14(4):934. <https://doi.org/10.3390/cancers14040934>
 50. Yin C, Cao Y, Sun P, Zhang H, Li Z, Xu Y, Sun H. Molecular Subtyping of Cancer Based on Robust Graph Neural Network and Multi-Omics Data Integration. *Frontiers in Genetics*. 2022;13:884028. <https://doi.org/10.3389/>

fgene.2022.884028

51. Choi JM, Chae H. moBRCA-net: a breast cancer subtype classification framework based on multi-omics attention neural networks. *BMC bioinformatics*. 2023 04 26;24(1):169. <https://doi.org/10.1186/s12859-023-05273-5>
52. Allogmani AS, Mohamed RM, Al-Shibly NM, Ragab M. Enhanced cervical precancerous lesions detection and classification using Archimedes Optimization Algorithm with transfer learning. *Scientific Reports*. 2024 05 27;14(1):12076. <https://doi.org/10.1038/s41598-024-62773-x>
53. Magdy Amin M, Ismail AS, Shaheen ME. Multimodal Non-Small Cell Lung Cancer Classification Using Convolutional Neural Networks. *IEEE Access*. 2024;12:134770-134778. <https://doi.org/10.1109/ACCESS.2024.3461878>
54. Kesimoglu ZN, Bozdog S. SUPREME: A cancer subtype prediction methodology integrating multiomics data using Graph Convolutional Neural Network. *bioRxiv*. 2022 08 05;. <https://doi.org/10.1101/2022.08.03.502682>
55. Basaad A, Basurra S, Vakaj E, Aleskandarany M, Abdelsamea MM. GraphX-Net: A Graph Neural Network-Based Shapley Values for Predicting Breast Cancer Occurrence. *IEEE Access*. 2024;12:93993-94007.
56. Geeitha S, Prabha KR, Cho J, Easwaramoorthy SV. Bidirectional recurrent neural network approach for predicting cervical cancer recurrence and survival. *Scientific Reports*. 2024;14(1):31641.
57. Ashayeri H, Sobhi N, Plawiak P, Pedrammehr S, Alizadehsani R, Jafarizadeh A. Transfer learning in cancer genetics, mutation detection, gene expression analysis, and syndrome recognition. *Cancers*. 2024;16(11):2138. <https://doi.org/10.3390/cancers16112138>
58. Tabassum F, Islam S, Rizwan S, Sobhan M, Ahmed T, Ahmed S, Chowdhury TM. Precision Cancer Classification and Biomarker Identification from mRNA Gene Expression via Dimensionality Reduction and Explainable AI. *arXiv preprint arXiv*. 2024;2410.07260.
59. Franchini M, Pellicchia S, Viscido G, Gambardella G. Single-cell gene set enrichment analysis and transfer learning for functional annotation of scRNA-seq data. *NAR genomics and bioinformatics*. 2023 03;5(1):lqad024. <https://doi.org/10.1093/nargab/lqad024>
60. Ming W, Li F, Zhu Y, Bai Y, Gu W, Liu Y, Sun X, Liu X, Liu H. Predicting hormone receptors and PAM50 subtypes of breast cancer from multi-scale lesion images of DCE-MRI with transfer learning technique. *Computers in Biology and Medicine*. 2022 Nov;150:106147. <https://doi.org/10.1016/j.compbio.2022.106147>
61. Pan L, Gao Q, Wei K, Yu Y, Qin G, Wang T. A robust transfer learning approach for high-dimensional linear regression to support integration of multi-source gene expression data. *PLoS computational biology*. 2025 01;21(1):e1012739. <https://doi.org/10.1371/journal.pcbi.1012739>
62. Samee NA, Mahmoud NF, Atteia G, Abdallah HA, Alabdulhafith M, Al-Gaashani MSAM, Ahmad S, Muthanna MSA. Classification Framework for Medical Diagnosis of Brain Tumor with an Effective Hybrid Transfer Learning Model. *Diagnostics (Basel, Switzerland)*. 2022 Oct 20;12(10):2541. <https://doi.org/10.3390/diagnostics12102541>
63. Muhammad B, Özkaynak F, Varol A, Tuncer T. A novel deep feature extraction engineering for subtypes of breast cancer diagnosis: A transfer learning approach. In 2022 10th International Symposium on Digital Forensics and Security (ISDFS) (pp. 1-7). IEEE. 2022.
64. Zhang T, Hasib MM, Chiu Y, Han Z, Jin Y, Flores M, Chen Y, Huang Y. Transformer for Gene Expression Modeling (T-GEM): An Interpretable Deep Learning Model for Gene Expression-Based Phenotype Predictions. *Cancers*. 2022 09 29;14(19):4763. <https://doi.org/10.3390/cancers14194763>
65. Wang YM, Sun Y, Wang B, Wu Z, He XY, Zhao Y. Transfer learning for clustering single-cell RNA-seq data crossing-species and batch, case on uterine fibroids. *Briefings in Bioinformatics*. 2023 Nov 22;25(1):bbad426. <https://doi.org/10.1093/bib/bbad426>
66. Attallah O. CerCan-Net: Cervical cancer classification model via multi-layer feature ensembles of lightweight CNNs and transfer learning. *Expert Syst. Appl.*. 2023 Nov 01;229(PB). <https://doi.org/10.1016/j.eswa.2023.120624>
67. Srikanthamurthy MM, Rallabandi VPS, Dudekula DB, Natarajan S, Park J. Classification of benign and malignant subtypes of breast cancer histopathology imaging using hybrid CNN-LSTM based transfer learning. *BMC medical imaging*. 2023 01 30;23(1):19. <https://doi.org/10.1186/s12880-023-00964-0>



This work is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License.