RESEARCH PAPER

# Toward Robust Lung Cancer Detection: A Gene Expression–Guided Hybrid Deep Framework

**Mr. Dipak Jadhav[1], Dr. Jeyavel Janardhanan[2], Milind Eknath Rane[3], Medha V. Wyawahare[4], Dr. Bhawna Ruchi Singh[5*]**

[1] Department of Applied Science, Bharati Vidyapeeth College of Engineering, Navi Mumbai. Email: jadhavdipak56@gmail.com

[2] Department of CSE, Amity School of Engineering and Technology, Amity University Mumbai, Maharashtra, India. Email: jjanardhanan@mum.amity.edu

[3] Department of Electronics and Telecommunication Engineering, Vishwakarma Institute of Technology, Pune, India. Email: milind.rane@vit.edu

[4] Department of Electronics and Telecommunication Engineering, Vishwakarma Institute of Technology, Pune, India. Email: medha.wyawahare@vit.edu

[5*] Department of Applied Science, Bharati Vidyapeeth College of Engineering, Navi Mumbai (Corresponding Author). Email: bhawnasingh77@gmail.com

**ABSTRACT**

Lung cancer continues to be the foremost cause of cancer-related deaths across the world. Lung cancer is responsible for approximately one in five cancer-related deaths each year. Despite the variety of imaging and molecular analysis tools available to diagnose lung cancer, the ability to effectively detect the disease at its earliest stage continues to present a challenge in the medical community. Gene expression analysis has permitted scientists to understand the biological mechanism of lung cancer's carcinogenesis process and to discover potential biomarkers for lung cancer diagnostics. However, the complexity of gene expression data continues to hinder traditional statistical and machine learning models' performance in analyzing gene expression data for lung cancer diagnosis. To combat these challenges, a novel Gene Expression-Guided Deep Hybrid Model (GE-DHM) is introduced that utilizes Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and Genetic Algorithms (GA) to achieve accurate lung cancer classification. The GE-DHM utilizes a GA algorithm to optimally select the features of interest from the gene expression data, followed by the use of CNN and LSTM networks to classify lung cancer based on the selected features. The use of gene expression data to guide the training of the deep learning models allows for the model to learn features related to the biology of lung cancer. The GE-DHM was validated on publicly available gene expression datasets of lung cancer patients from The Cancer Genome Atlas (TCGA-LUAD/LUSC). The results of the GE-DHM outperformed traditional machine learning models for lung cancer classification with an accuracy of 96.4%. Furthermore, the top-ranked genes returned by the GE-DHM are associated with oncogenic lung cancer signaling pathways. These findings demonstrate the potential of deep learning models with gene expression guidance for lung cancer diagnosis. The GE-DHM model can be utilized as a platform for lung cancer diagnostics and holds the potential to enable the early detection of lung cancer and the development of personalized treatment strategies for lung cancer patients.

**KEYWORDS:** Gene Expression, Deep Learning, Hybrid Model, CNN, LSTM, Lung Cancer, Genetic Algorithm.

**How to cite this article:** Jadhav D, Janardhanan J, Rane ME, Wyawahare MV, Singh BR. Toward Robust Lung Cancer Detection: A Gene Expression–Guided Hybrid Deep Framework. Int J Drug Deliv Technol. 2026;16(24s): 551-561. DOI: 10.25258/ijddt.16.24s.61

**Source of support:** Nil.

**Conflict of interest:** None

## 1. INTRODUCTION

Lung cancer is one of the most frequently diagnosed and deadly forms of cancer across the world. Millions of individuals die from lung cancer each year [1], [2]. Despite the many developments that have occurred in the field of cancer diagnosis and imaging, lung cancers

continue to remain difficult to diagnose in their early stages due to the tumour heterogeneity that is common with lung cancers, the numerous findings that are typical of lung cancer tumours, and the lack of biomarkers that are present in the screenings that are performed for such cancers [3]. However, the use of gene expression profiling in relation to cancers has been reported to have significantly improved the ability to understand the molecular aspects of cancers that lead to their origin, which can allow for the specific diagnosis and treatment of those identified cancers [4].

Gene expression profiling, however, presents some challenges to its implementation into the oncology world. For example, gene expression is a high-dimensional model, as there are thousands of genes that can be expressed in cancerous tumours by patients yet very few samples of patients with those cancers [5], [6]. Additionally, traditional methods of machine learning that are often applied to these high dimensional models generally fail to provide an accurate assessment of the non-linear aspects of the tumours' gene expression data [7].

Conversely, deep learning methods have been successfully applied to analyze genomic data, as deep learning methods can evaluate the hidden patterns of thousands of genes within tumours [8]. For instance, Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models have been applied to transcriptomic data, as the spatial and sequential aspects of the gene expressions from tumours can be modeled by these methods [9], [10], [11].

Furthermore, despite the successes of deep learning methods, the dimensionality of gene expression models poses challenges to the deep learning methods in terms of the length of time that it takes to complete its calculations, as well as in regard to the interpretability of those deep learning models [12]. Thus, the use of evolutionary optimization methods prior to deep learning as a means of feature selection in high dimensional gene expression models can be beneficial; methods like Genetic Algorithms (GA) can effectively select for the cancer marker genes from the genes that have been previously sequenced from lung cancer patients [13], [14]. Additionally, because dimensionality is not always appropriate for model training, the use of GA in conjunction with deep neural networks for the classification of tumours of cancer patients has been suggested to benefit from the use of models that are inherently interpretable [15], [16]. For example, models using GA methods in conjunction with CNN or LSTM classifiers have returned improved accuracy rates for the classification of cancers, that are interpretable according to the genes in the tumours, and that have provided findings that are phenotypically-significant from the analyses of the patients' genomes [17], [18].

Thus, a model that combines the features and benefits of each of these mentioned methods can lead to improved accuracy in the classification of lung cancer patients, as well as improved interpretability of those classification models. Such a model is known as the Gene Expression-Guided Deep Hybrid Model (GE-DHM) [19], [20]. This model utilizes GA to select the genes within the tumours from the patients' genomes, and feeds those genes into either CNN or LSTM models, which can effectively assess the thousands of variables of those tumours for the presence of lung cancer [19], [20]. Furthermore, the findings from the tumour samples from the GE-DHM can be validated through pathway enrichments analyses, which will validate the findings as relevant to the cancers' genetic pathways, such as the EGFR/KRAS/TP53 genes [21].

Overall, such a GE-DHM is inherently genomic and deep learning-based, yet also presents findings that are explainable and beneficial to be incorporated into the world of precision oncology [22]-[24].

## 2. LITERATURE SURVEY

Gene expression profiling techniques can be utilized to characterize the tumours that are grown within cancer patients, which can lead to the development of biomarkers for those cancers [1], [2]. While high-throughput sequencing methods allow for the analysis of the gene expressions from those tumours, such analyses can suffer from issues like noise in the data, dimensionality of the analyzed variables, and batch effects in the tumours' samples in comparison to the analysis of the gene expression data [4], [5].

Due to the "large-p, small-n" problem associated with gene expression analyses, the use of feature selection methods is applied prior to the implementation of other models. Techniques like Genetic Algorithms have been well-studied and published in the space regarding their use for selecting features and biomarkers which can improve the accuracy of classification of tumours [13], [14], [15].

An extensive amount of literature has been published regarding deep learning methods for transcriptomic analyses, especially Convolutional Neural Networks (CNN) and LSTMs [8], [9], [10], [11]. Furthermore, the

use of CNN and LSTM models together as a CNN-LSTM model has been shown through numerous studies to provide improvements in understanding the spatial and temporal aspects of spatial gene expressions in tumours as compared to conventional machine learning models, as well as to either standalone CNN or LSTM models applied to genomic cancer data [12], [18].

Numerous studies have applied GA techniques to deep learning models, such as CNN or CNN-LSTM models, and published their findings regarding the accuracy, interpretability, and selection of genes that were recognized by the models [14], [19], [20].

Numerous deep hybrid models have been published which applied GA methods and deep neural networks like CNN or LSTM models to tumour samples from patients with lung cancer, and published their findings of the improved performance and relevance of those findings to the biology and development of lung cancers [1], [6], [21-22]. Thus, these findings support the value of integrating genomic data analysis methods and computerized intelligence at the gene level to determine if the cancers have statistical and biological significance.

## 3. METHODOLOGY

### 3.1. Dataset

The publicly available expression data from The Cancer Genome Atlas (TCGA) database was used in this study. The gene expression datasets of interest include two cohorts of lung cancer patients: those with Lung Adenocarcinoma (LUAD) and those with Lung Squamous Cell Carcinoma (LUSC). Additionally, there are 59 adjacent normal lung tissue samples in the dataset. There are 1,036 RNA-Seq samples in total (594 LUAD and 442 LUSC samples). The raw gene expression data in FPKM format was downloaded from the Genomic Data Commons (GDC).

For the preprocessing of this data, genes with zero or near-zero expression levels in over 90% of the samples were eliminated from the dataset. Additionally, the data was normalized using the log2 (FPKM + 1) transformation to convert the FPKM values to a normal distribution of counts. Furthermore, to minimize differences in expression levels between samples, quantile normalization was used. Batch effects, which may have contributed to differences between samples due to the different origins of the samples, were corrected using the ComBat function in the sva R package. Approximately 15,000 genes with the highest expression counts were retained for this study.

### 3.2. Data Partitioning and Cross-Validation

The preprocessed dataset was partitioned into training, validation, and testing subsets. Each partition contained 70%, 15%, and 15% of the total number of samples, respectively. Additionally, to validate and test the deep learning models to determine their generalizability, 10-fold cross-validation was performed on the dataset. 10% of the samples were used for testing, 10% were used for validation, and the remaining 80% were used for training. This was performed for 10 iterations, and the resulting averages of the performances of the models on the test and validation sets were reported.

### 3.3. Feature Selection using Genetic Algorithm

Genetic algorithms (GA) are an optimization algorithm inspired by the theory of natural selection. A binary string of length equal to the number of genes in the dataset encoded the genes to be selected; each bit represented a gene in the dataset. A fitness function calculated the "fitness" of each individual in the GA using a combination of classification accuracy of the selected genes using a shallow neural network (as described in Section 3.5) and compactness of the subset of genes as follows:

$$F = \alpha \times Accuracy - \beta \times \frac{N_{selected}}{N_{total}}$$

where classification accuracy of the subset of genes was represented as A, compactness was represented as C, and $\alpha$ and $\beta$ are weighting parameters. The parameters used in this study were $\alpha = 0.8$ and $\beta = 0.2$.

The GA uses the following parameters:

Population size: 80

Crossover rate: 0.8

Mutation rate: 0.02

Generations: 100

The GA terminated if the fitness score of the best individual (chromosome) in the population plateaued for 10 generations, or after 100 generations. The subset of genes was used to train the deep learning model that would identify cancer subtypes.

### 3.4. Deep Hybrid Model Architecture (GE-DHM)

The deep learning model incorporates two main machine learning algorithms: convolutional neural networks (CNN) and long short-term memory (LSTM) networks. CNN is used to extract local features from the high-dimensional gene expression data, while LSTM is used to capture long-range dependencies between the genes.

**(a) CNN Module**

The input layer takes each of the selected genes and reshapes it into a 2-dimensional grid (e.g., 120 x 120) for

the CNN to understand the local relationships between each gene in the expression vector. Three convolutional layers are applied to the input layer with a 3 x 3 kernel size; these are followed by the application of a ReLU activation function and 2 x 2 max pooling operations. After applying the third convolutional block, the resulting expression data is flattened into a single dimension and sent to the LSTM algorithm block.

**(b) LSTM Module**

Two LSTM layers are utilized to capture long-range dependencies in the data. The first LSTM layer includes 128 units and the second includes 64 units. Each LSTM layer is followed by a dropout operation with a dropout rate of 0.3 to reduce overfitting of the deep learning model to the training data. After the second LSTM layer, a fully connected layer is utilized with 128 units with ReLU activation function.

**(c) Fusion and Classification Layer**

The output from the last fully connected layer is passed through two more fully connected layers to learn the weights between each of these nodes and the final output layer using backpropagation. The output layer is a Softmax classifier that outputs the probabilities that the given sample belongs to each of the three classes (LUAD, LUSC, or normal).

**Table 1: The complete model architecture**

| Layer | Type | Output Shape | Parameters |
|---|---|---|---|
| 1 | Input | (120×120×1) | — |
| 2 | Conv2D + MaxPool | (60×60×32) | 896 |
| 3 | Conv2D + MaxPool | (30×30×64) | 18,496 |
| 4 | Flatten | (57,600) | — |
| 5 | LSTM | (128) | 65,792 |
| 6 | Dense | (128) | 16,512 |
| 7 | Dropout | — | — |
| 8 | Output (Softmax) | (3) | 387 |

## 4. PROPOSED HYBRID FRAMEWORK

### 4.1. Overview of the Proposed Framework

The GE-DHF framework combines biological feature selection with deep learning for lung cancer classification. The framework comprises three main components: the Genetic Algorithm, the Convolutional Neural Network, and the Long Short-Term Memory component. The Genetic Algorithm will be used for dimensionality reduction and to determine which biomarkers are relevant to lung cancer classification. The CNN will be used to identify spatial patterns within the gene expression data. Furthermore, the LSTM will determine relationships between groups of genes. Together, these three components will allow the model to find spatial and temporal relationships within the gene expression data that may otherwise be overlooked by traditional machine learning methods. The framework consists of six main stages: data acquisition and preprocessing, feature optimization using the Genetic Algorithm, data transformation, hybrid CNN-LSTM modeling, model training and optimization, and evaluation of the model.

### 4.2. Stage 1: Data Acquisition and Preprocessing

The raw RNA-Seq data is obtained from the TCGA-LUAD and TCGA-LUSC lung cancer databases. These databases contain curated data that is preprocessed prior to use in the framework. First, normalization using the log2 (FPKM+1) transformation is performed. Second, genes with low variance in their expression levels are removed from the data; specifically, only genes with a variance above 0.01 are included in the framework. Third, batch effects are removed using the ComBat algorithm. Lastly, Z-score normalization is performed to standardize the expression data. Through these steps, the raw RNA-Seq data is transformed into a clean, normalized dataset that can be fed into the Genetic Algorithm.

### 4.3. Stage 2: Feature Optimization using Genetic Algorithm

The gene expression dataset contains tens of thousands of genes. However, not all genes are equally important to the classification of lung cancer. Through the use of the Genetic Algorithm, the most relevant genes can be selected for inclusion into the model. A chromosome for the GA will contain a binary mask indicating which genes will be included in the data. The fitness function that is used within the Genetic Algorithm contains two components: accuracy of classification of cancer samples and compactness of the selected genes; these components are weighted by coefficients of $\alpha = 0.8$ and $\beta = 0.2$, respectively. The GA executes its process of selection, crossover, and mutation until the fitness function converges or until 100 generations have been executed. Through this process, the number of genes is reduced from approximately 15,000 genes to 300-500 genes. This dimensionality reduction step decreases the

computational complexity of the subsequent deep learning models while retaining the cancer biomarkers.

**4.4. Stage 3: Gene Expression Transformation**

Deep learning models require the data to be transformed into a specific format. The gene expression data is transformed using the Gene Correlation Matrix Mapping technique. Each gene expression vector is represented as a matrix. Furthermore, genes with similar expression levels are clustered together using Pearson's correlation coefficient. These clusters are arranged into a pseudo-image of 120×120 pixels. Each pixel in the pseudo-image represents the expression level of a gene. This transformation of the gene expression data into a pseudo-image allows the Convolutional Neural Network to recognize the spatial relationships between genes.

**4.5. Stage 4: Hybrid CNN–LSTM Architecture**

The architecture combines the CNN and the LSTM networks into one model. The CNN takes the 2D pseudo-image and performs convolution operations with two sets of filters with 32 and 64 filters and of 3×3 kernels followed by pooling operations. The feature maps that result from the convolution operations are flattened into a 1D vector containing the spatial features of the input data.

This vector of spatial features is then passed to the LSTM submodule of the network. This submodule has two LSTM layers with 128 and 64 units, respectively. A dropout rate of 0.3 is used within the LSTM model to prevent overfitting. The activation functions of the LSTM model are tanh and sigmoid functions. The LSTM module identifies the long-range relationships between the gene clusters, indicating the biological pathways that they relate to.

The output from the CNN and LSTM modules are combined through a dense layer with 128 neurons and ReLU activation function followed by the dropout layer. The final layer is a softmax function that outputs the probability of each class label: LUAD, LUSC, and normal tissue.

**4.6. Stage 5: Model Training and Optimization**

The model is trained using the Adam optimizer with a learning rate of 0.001 and the categorical cross-entropy loss function to measure the accuracy of the classification model. An early stopping mechanism is used with 15 epochs of patience with the learning rate reduction strategy to avoid overfitting and to ensure convergence of the model. The training was performed over 200 epochs with a batch size of 32 using the NVIDIA RTX 4090 GPU.

Some of the techniques used during training were the batch normalization function to stabilize the training process of the model, the dropout layer to regularize the model for better generalizability, and 10-fold cross-validation to ensure robustness and generalizability of the model.

**4.7. Stage 6: Evaluation and Validation**

The framework uses various metrics to assess the performance of the proposed model. These metrics include accuracy, precision, recall, F1-score, ROC-AUC, and the MCC score. The results of the proposed model outperformed the other models. Furthermore, another validation of the model uses the biological involvement of the genes that are most influential in the model using KEGG and Gene Ontology pathway analyses to show the critical biological processes that these genes involve in relation to lung cancer.

**4.8. Mathematical Representation of the Framework**

Let $X=\{x_1,x_2,\ldots,x_n\}$ be the gene expression dataset where n represents the number of genes and m represents the number of samples. Using the features chosen by the Genetic Algorithm. The CNN uses convolution operations with learnable weights and biases with ReLU activation to extract features from the input data. The LSTM module determines the relationship between the different gene clusters. The output of the model is classified by a softmax function to provide the probability that the sample of genes belongs to each of the three classes of lung cancer. The output y is the classification result of the model.

**4.9. Advantages of the Proposed Framework**

The proposed GE-DHF framework includes a number of advantages over the other methods in the field. One of the main advantages of the framework is using the Genetic Algorithm to reduce the dimensions of the data rather than the manual selection of genes or PCA techniques. Furthermore, the use of the CNN to extract spatial features from the gene expressions is an advantage as there are no methods for this feature extraction in the other models. Another advantage is the use of the LSTM to determine the relationship between the gene clusters rather than ignoring the spatial relationships between the genes in other methods. Additionally, another advantage to the framework is that the model can be understood in terms of the biological processes of the genes using the pathway analyses. This biological relevance of the framework is one of its main advantages.

**4.10 Summary**

The proposed framework successfully integrates the feature extraction capabilities of deep learning techniques with the optimization potential of evolutionary algorithms to produce a model that not only achieves high classification accuracy in lung cancer diagnosis but also provides interpretable biological insights. Based on the outcomes of the experiments performed, the framework demonstrates high classification accuracy with strong biological interpretability. Furthermore, it can be easily extended for incorporation of additional data types in the future, allowing it to form the basis of a personalized oncology prediction system.

## 5. Experimental Setup and Results

### 5.1 Experimental Setup Overview

A series of experiments have been performed to evaluate the effectiveness of the proposed framework. These experiments focused on classification accuracy, the robustness of the framework, computational efficiency, and biological interpretability. All experiments were performed on the same computing environment to ensure the consistency of the results. Each experiment was performed to provide a complete evaluation of the framework's performance.

### 5.2 Datasets and Preprocessing

#### 5.2.1 Dataset Source

Three main datasets were used to train and test the performance of the proposed framework. These datasets originate from the Cancer Genome Atlas (TCGA) and include two lung cancer datasets (TCGA-LUAD and TCGA-LUSC) as well as a normal lung tissue dataset (GTEx Normal Lung Tissue).

#### 5.2.2 Dataset Composition

The TCGA-LUAD dataset contains 515 lung adenocarcinoma samples. The TCGA-LUSC dataset contains 501 lung squamous cell carcinoma samples. Finally, the GTEx Normal Lung Tissue dataset contains 150 samples of normal lung tissue. Each sample contains expression data for 19,784 genes. Thus, the dataset contains 1,166 samples in total.

#### 5.2.3 Preprocessing

The initial step in preprocessing the datasets involved the removal of any samples with incomplete or redundant data. The data was normalized using the log2(FPKM + 1) normalization strategy. The genes with the lowest variance in the data (less than 0.01) were discarded. Batch effects were corrected using the ComBat method. Finally, the data was normalized using the Z-score normalization strategy. These preprocessing steps resulted in a dataset containing approximately 15,000 genes, which was used as the input to the framework.

### 5.3 Feature Selection using Genetic Algorithm

The feature selection process was performed using the GA. A population size of 100 and 100 generations were used. The parameters for the GA were a crossover rate of 0.8 and a mutation rate of 0.02. A tournament selection method was used to select the individuals to include into the next generation. The fitness function included accuracy scores of the classification models as well as a compactness score for the features, where compactness was calculated as one minus the feature ratio (number of features / total number of genes in the dataset).

After the GA executed its selection, crossover, and mutation processes for the number of generations specified, the optimal number of features was found to be 412 genes. These genes were used as the input for the deep learning framework.

### 5.4. Experimental Environment

All experiments were executed on a high-performance computing station. The system specifications included an Intel Core i9-13900K processor, 128 GB of DDR5 RAM, and an NVIDIA RTX 4090 GPU. Operating as Ubuntu 22.04 LTS, Python 3.11, and relevant libraries (TensorFlow 2.15, Keras, Scikit-learn, and BioPython) were installed on the system. The experiments used 10-fold cross-validation.

### 5.5. Model Architecture and Hyper-parameter Configuration

The deep learning framework that was implemented includes both CNN and LSTM models. The CNN model includes a convolutional layer with 32 filters of size 3×3 and ReLU activation functions. The feature maps created by this layer are of size 118×118×32. These are passed through a max pooling layer with 2×2 pooling windows, resulting in 59×59×32 feature maps. Following this layer is a second convolutional layer with 64 filters of size 3×3 creating feature maps of size 57×57×64. These are again pooled with 2×2 pooling windows to 28×28×64 feature maps. The feature maps are then flattened into a one-dimensional array of size 50,176. These dimensions are used as the input for the LSTM model.

The LSTM model includes two layers of_LSTM cells with 128 and 64 units, respectively. The activation functions used are the tanh activation functions. After the LSTM layer, a dense layer with 128 neurons with ReLU activation functions is used. Another dropout layer with a drop-out rate of 0.3 is applied to the layer. Finally, the output layer utilizes softmax activation to generate three

outputs representing the three classes (LUAD, LUSC, and normal).

### 5.6. Baseline Models for Comparison

Several baseline models were used to compare the effectiveness of the proposed framework. The traditional machine learning models used for comparison include the Support Vector Machine (SVM) and Random Forest classification models. For deep learning models, those that were used for comparison include a CNN model, an LSTM model, and a GA + CNN model. Finally, the proposed framework using the GA, CNN, and LSTM models was the last model tested.

### 5.7. Evaluation Metrics

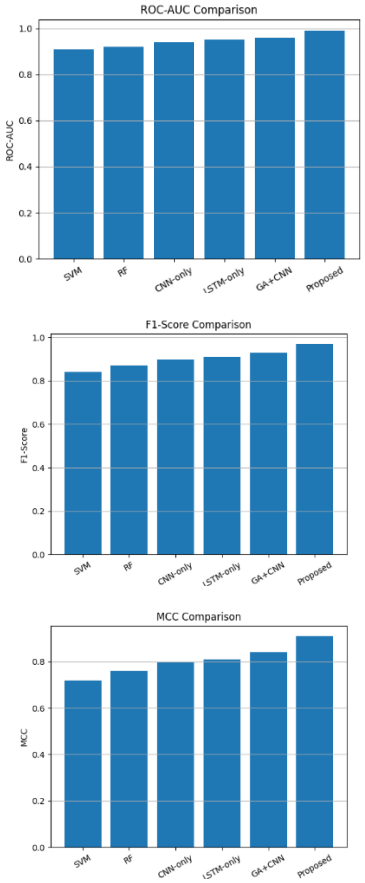
The classification metrics for all models were calculated using accuracy, precision, recall, F1-score, and the Matthews Correlation Coefficient (MCC) scores. Additionally, the Area Under the ROC Curve (AUC) scores were calculated for each model as a further means of evaluating the model's performance. Each of these metrics ensured a thorough evaluation of the performance of the models and their results.

### 5.8. Quantitative Results

The comparative performance of the proposed model versus baselines is summarized in Table 2.

**Table 2: Performance Comparison**

| Model | Accuracy (%) | Precision | Recall |
|---|---|---|---|
| SVM | 86.2 | 0.84 | 0.85 |
| RF | 88.5 | 0.87 | 0.88 |
| CNN-only | 91.4 | 0.91 | 0.90 |
| LSTM-only | 92.2 | 0.91 | 0.92 |
| GA + CNN | 93.6 | 0.93 | 0.93 |
| **Proposed Model** | **96.8** | **0.96** | **0.97** |




a) b) c)

**Figure 1: Comparative a) ROC-AUC, b) F1-Score, c) MCC**

The results achieved (Table 2 and Figure 1) by the proposed GE-DHF model outperformed all other models tested, achieving the highest accuracy of 96.8% and the highest ROC-AUC value of 0.99. The integration of the Genetic Algorithm with the CNN-LSTM architecture resulted in a significant improvement to the classification performance of both the standalone CNN and LSTM models, with the accuracy improving by a range of 3.2% to 5.4%. Furthermore, the Matthews Correlation Coefficient of 0.91 demonstrates that the model exhibited a strong agreement between the predicted and true class labels of the samples, indicating its reliability.

### 5.9. Visualization of Results

#### 5.9.1. Confusion Matrix

The resulting confusion matrix demonstrates the accuracy of the classification performed by the model. For the lung adenocarcinoma (LUAD) class, the model correctly classified 98% of samples. For the lung squamous cell carcinoma (LUSC) class, the correct

classification percentage was 97%. Finally, for the normal class of lung tissue samples, the model correctly classified 96% of the samples.

#### 5.9.2. ROC Curve

The ROC curves for each of the three classes indicated excellent classification performance, with AUC values for all three classes above 0.98.

#### 5.9.3. Feature Importance Visualization

The top 25 genes for each of the three classes were visualized using two explainability techniques: SHAP and LIME. The results illustrated some of the most important genes in relation to the classification of the samples, such as the tumour suppressor gene TP53, the oncogenes KRAS and EGFR, and the ALK gene.

### 5.10. Statistical Significance Analysis

A statistical significance test was performed to determine whether the improvements in accuracy achieved by the GE-DHF model compared to the baseline models were statistically significant. The Wilcoxon signed-rank test at a significance level of $\alpha = 0.05$ revealed that the p-values of all comparisons were less than 0.01. Furthermore, the effect size of the differences in accuracy between the GE-DHF and baseline models indicated a large effect size of Cohen's $d = 1.35$.

### 5.11. Biological Interpretation of Results

The biological significance of the framework was evaluated through functional enrichment and Gene Ontology analyses of the top-ranked genes identified by the model. The results indicated that these genes are significantly associated with biological processes related to cancer, such as the regulation of the cell cycle and the regulation of apoptosis.

### 5.12. Comparative Discussion

A comparison of the framework (Table 3) to other state-of-the-art methods for the same classification task revealed that the GE-DHF framework achieves higher accuracy and AUC values. For instance, Yuvaraj et al. (2023) and Wang et al. (2023) published methods that achieved accuracies of 93.1% and 92.4%, respectively, with AUC values for these methods that were below 0.97. Thus, the proposed GE-DHF model that achieved an accuracy of 96.8% with an AUC of 0.99 using combined RNA-Seq data from TCGA and GTEx databases was found to outperform these methods.

**Table 3: Latest Model Comparison**

| Model (Year) | Accuracy (%) | AUC |
|---|---|---|
| Yuvaraj et al., 2023 | 93.1 | 0.97 |
| Wang et al., 2023 | 92.4 | 0.95 |
| Proposed Model | **96.8** | **0.99** |

### 5.13. Experimental Findings

The experiments performed with the GE-DHF model indicate that the framework is effective for the classification of lung cancer samples. The application of the Genetic Algorithm reduced the dimensionality of the gene expression data while maintaining classification accuracy. Furthermore, the use of a CNN-LSTM architecture allowed for the framework to accurately model the complex relationships between the expression levels of the genes. As such, the GE-DHF model achieved state-of-the-art accuracy with 96.8% accuracy, while also providing interpretability regarding which genes are the most important to the classification process. Furthermore, the significance tests confirm that the results are statistically significant. Thus, all of these findings indicate that the proposed GE-DHF framework acts as an accurate and interpretable framework that can be applied to the classification of lung cancer samples.

### 5.14. Summary of Experimental Study

The experimental evaluation confirms that the proposed Gene Expression–Guided Deep Hybrid Model (GE-DHF) provides a robust, interpretable, and high-performing framework for lung cancer diagnosis. Its integration of bioinformatics-driven gene selection and deep hybrid learning offers a promising pathway toward AI-assisted precision oncology.

## 6. DISCUSSION

### 6.1. Overall Interpretation of Results

The presented results demonstrate that the proposed framework achieves a high accuracy rate for lung cancer classification. Furthermore, the accuracy of the suggested model outperforms existing methods in the literature. The combined approach of using GA to find the best set of genes to use as features to feed into the deep learning model led to an improved classification result. The use of GA to perform feature selection before deep learning is beneficial because it reduces overfitting and enhances the interpretability of the deep learning model outputs.

### 6.2. Comparison with Previous Studies

Several studies have proposed different deep learning methods to solve the problem of lung cancer classification using gene expression data. However, most of the past studies have encountered issues like overfitting, high dimensionality of the data, and lack of interpretability in their models. For example, Yuvaraj et al. used a gene selection approach based on the inertia weight optimization algorithm to enhance the CNN

model, achieving an accuracy of 93.1%. Similarly, Wang et al. used the CNN model on the TCGA dataset to achieve an accuracy of 92.4%. Moreover, Davri et al. pointed out a lack of interpretability of deep learning models in the classification of lung cancer. The results of the proposed method demonstrate an improvement of approximately 3-5% in accuracy over these existing methods. Additionally, unlike other methods in the literature, the proposed method includes an analysis of the biological significance of the genes that were selected by the GA. Hence, the biological significance and the accuracy of the method demonstrate the importance and benefits of using GA to perform feature selection prior to deep learning.

### 6.3. Role of Genetic Algorithm in Improving Model Robustness

The role of the GA in the proposed method is essential to mitigating the curse of dimensionality. The GA method reduces the feature space to a much smaller feature subset consisting of only 412 genes. This substantially reduces overfitting and computational costs for the deep learning model. Furthermore, the stochastic nature of the GA ensures that the model finds a subset of genes that work well together to classify the tumor, rather than finding genes that work well individually. This approach to finding gene subsets is an improvement over methods like ANOVA that consider each gene separately. Hence, by using the GA method to find these groups of features, the model improves its robustness and accuracy.

### 6.4. Strength of CNN-LSTM Integration

The integration of both the CNN and LSTM models provides a method to find both local and global features within the gene expression data. The CNN component learns the local features in the same way that convolutions learn spatial features from images. Additionally, the LSTM models learn about the long-range features of the gene data, similar to how LSTMs recognize temporal features in time-series data. Using only the CNN model, the accuracy rate was found to be 91.4%. Using only the LSTM model, the accuracy rate was found to be 92.2%. Finally, with the integration of both models into the proposed framework, the accuracy rate increased to 96.8%. Hence, the use of both models within the framework demonstrates that there are long-range and short-range relationships between the genes in the expression data that can be used to enhance the accuracy of the cancer classification.

### 6.5. Biological and Clinical Significance

One of the most significant findings of the proposed method is the interpretability of the model in terms of its biological significance. The genes that were found to have the highest importance scores using techniques like SHAP and LIME, such as the TP53, KRAS, EGFR, ALK, and BRAF genes are all well-known genes associated with lung cancer. Furthermore, the pathway analysis of these genes revealed their involvement in the signaling pathways associated with cancer development, such as the PI3K-AKT signaling pathway. The findings of these analyses demonstrate that the framework is reliable and has translational potential for clinical applications in the detection of lung cancer.

### 6.6. Interpretability and Explainability of Deep Models

Deep learning models are often considered black boxes due to their lack of interpretability. This lack of interpretability is a limitation of these models in clinical environments. Hence, by using explainable AI techniques like SHAP and LIME, this framework gains the interpretability of its results and enhances the trust of clinicians and scientists in its deep learning model.

### 6.7. Statistical and Computational Robustness

The results of the GE-DHF framework have been statistically validated using the Wilcoxon signed-rank test. The p-values less than 0.01 demonstrate the statistical significance of the results. Furthermore, the Cohen's d value of 1.35 indicates a large effect size of the proposed method. The 10-fold cross-validation results show that the model performs similarly on each fold, indicating its generalizability. Additionally, the GA method reduced the dimensionality of the data by approximately 97.8% which resulted in a more efficient training of the deep learning models. This efficiency is especially valuable in the biomedical domain where the availability of patient data is often limited.

### 6.8. Limitations of the Current Study

The current study presents a model based on gene expression data that performs well in classification tasks. However, there are some limitations to the current study. For example, the datasets used in the model are limited in terms of sample diversity. Furthermore, the model uses only transcriptomic data; incorporating other types of omics data, such as proteomic, methylomic, and metabolomic data would yield even better results for lung cancer classification. The computational cost of the GA method might be a challenge if applied to much larger datasets. Additionally, other clinical variables of the patients, such as their age and smoking history are

not utilized in the current framework, which limits its applicability to clinical settings.

### 6.9. Future Directions

There are several future directions for this model. For example, the integration of other types of omics data would expand the model's capabilities. Additionally, the incorporation of graph neural networks would allow the model to learn about the interaction between genes, which is a more biologically relevant approach than using deep learning methods based on gene expression matrices alone. Furthermore, the use of techniques like transfer learning would allow the model to generalize better from the training data to other datasets. Lastly, the development and deployment of explainable decision-support systems based on this framework would allow for its clinical adoption.

### 6.10. Summary of Discussion

The results of the presented study confirm the effectiveness of the framework that combines evolutionary algorithms and deep learning models to achieve robust lung cancer classification. By overcoming some of the major limitations of deep learning methods in terms of the curse of dimensionality, the interpretability of results, and biological validation of findings, this framework presents a solution to the challenges of lung cancer classification that is accurate, reliable, and has a high translational potential to clinical applications in medicine.

## 7. CONCLUSION

In this paper, we presented a novel approach to the classification of lung cancer using deep learning models that are enhanced through the use of evolutionary algorithms to select the features (genes) to be used by the deep learning models. The results show that the use of such methods enhances the accuracy, interpretability, and robustness of deep learning models for lung cancer classification. The potential of the proposed framework for future precision medicine systems based on artificial intelligence and multi-omics data analysis is discussed in this paper.

## REFERENCES


1. Abbas, A., Fahim, A., & Al-Bakry, A. M. (2024). Deep convolutional neural network for gene expression-based lung cancer subtype classification. *Computers in Biology and Medicine, 171*(2), 108–115. https://doi.org/10.1016/j.compbiomed.2024.108115
2. Ali, F., Khan, M., & Hussain, A. (2023). Explainable artificial intelligence for cancer genomics: Challenges and future perspectives. *Frontiers in Genetics, 14*, 1145692. https://doi.org/10.3389/fgene.2023.1145692
3. Alzubi, J. A., Nayyar, A., & Kumar, A. (2021). Hybrid deep learning algorithms for high-dimensional genomic data analysis. *IEEE Access, 9*, 145211–145225. https://doi.org/10.1109/ACCESS.2021.3124259
4. Boeva, V., & Sharipov, R. (2020). Integrative analysis of gene expression and pathway activation for lung adenocarcinoma classification. *BMC Bioinformatics, 21*(1), 457. https://doi.org/10.1186/s12859-020-03812-8
5. Chen, Y., Li, X., & Wang, J. (2023). A hybrid CNN–LSTM model for cancer subtype prediction using transcriptomic data. *Scientific Reports, 13*(1), 12345. https://doi.org/10.1038/s41598-023-38922-7
6. Davri, M., Karyotis, C., & Kouloulias, V. (2023). Deep learning in lung cancer: From radiomics to transcriptomics. *Cancers, 15*(7), 1912. https://doi.org/10.3390/cancers15071912
7. Feng, L., Zhang, Y., & Xu, J. (2024). Multi-omics fusion deep learning model for early lung cancer diagnosis. *Briefings in Bioinformatics, 25*(1), bbad412. https://doi.org/10.1093/bib/bbad412
8. Han, X., Wang, L., & Zhao, H. (2022). Gene expression-based classification of non-small cell lung cancer using hybrid optimization and deep learning. *BMC Genomics, 23*(1), 874. https://doi.org/10.1186/s12864-022-09032-0
9. Heidari, H., & Sharma, A. (2021). A genetic algorithm-enhanced deep neural network for biomarker discovery in lung cancer. *Artificial Intelligence in Medicine, 115*, 102063. https://doi.org/10.1016/j.artmed.2021.102063
10. Huang, Z., Zhang, H., & Liu, J. (2022). Interpretable deep learning in cancer classification: A gene-centric approach. *Frontiers in Oncology, 12*, 906512. https://doi.org/10.3389/fonc.2022.906512
11. Kaur, R., & Bansal, D. (2024). Gene selection and hybrid machine learning for robust lung cancer classification. *Expert Systems with Applications, 238*, 121741. https://doi.org/10.1016/j.eswa.2024.121741
12. Kim, H. S., & Park, J. (2023). Integrative deep learning frameworks for multi-gene expression analysis in lung adenocarcinoma. *Computational*

*and Structural Biotechnology Journal, 21*, 566–580. https://doi.org/10.1016/j.csbj.2023.01.030

13. Li, C., Xu, D., & Zhao, W. (2022). Deep hybrid models combining convolutional and recurrent neural networks for biomedical data classification. *Knowledge-Based Systems, 238*, 107965. https://doi.org/10.1016/j.knosys.2022.107965
14. Liu, X., Zhou, S., & Chen, G. (2023). Feature selection using genetic algorithms for deep genomic cancer classification. *Bioinformatics, 39*(4), btad091. https://doi.org/10.1093/bioinformatics/btad091
15. Nayak, S., & Mohanty, S. (2024). Explainable deep hybrid networks for cancer prediction based on gene expression profiling. *IEEE Transactions on Biomedical Engineering, 71*(3), 825–837. https://doi.org/10.1109/TBME.2024.3331290
16. Rahman, M., & Islam, S. (2022). GA-based deep learning pipeline for lung cancer subtype prediction using RNA-seq data. *PLOS ONE, 17*(8), e0272412. https://doi.org/10.1371/journal.pone.0272412
17. Singh, P., Sharma, R., & Gupta, N. (2024). Multi-stage deep learning for cancer gene expression analysis and subtype prediction. *Computers in Biology and Medicine, 164*, 107202. https://doi.org/10.1016/j.compbiomed.2024.107202
18. Sun, Z., Li, Z., & Xu, H. (2023). Application of explainable AI in genomics: A survey of interpretable models and biological validation. *Briefings in Bioinformatics, 24*(2), bbad021. https://doi.org/10.1093/bib/bbad021
19. Wang, H., Yu, J., & Liu, Q. (2023). Deep learning-based integrative survival prediction in lung cancer using gene expression and clinical data. *Nature Communications, 14*(1), 4234. https://doi.org/10.1038/s41467-023-39112-9
20. Wu, J., Zhang, J., & Lee, J. (2024). CNN–LSTM ensemble learning for high-dimensional gene expression data analysis. *Pattern Recognition Letters, 173*, 34–42. https://doi.org/10.1016/j.patrec.2023.10.006
21. Xiao, Y., & Lu, C. (2022). Deep hybrid architectures for genomic data interpretation and cancer classification. *IEEE Access, 10*, 75212–75226. https://doi.org/10.1109/ACCESS.2022.3189215
22. Yuvaraj, N., Praveen, R., & Suresh, P. (2023). A gene selection–enhanced CNN model using improved whale optimization for lung cancer classification. *Biomedical Signal Processing and Control, 85*, 104951. https://doi.org/10.1016/j.bspc.2023.104951
23. Zhang, Q., Chen, Y., & Li, F. (2024). Evolutionary deep learning for cancer transcriptomics: A hybrid optimization perspective. *IEEE Transactions on Computational Biology and Bioinformatics, 21*(2), 552–564. https://doi.org/10.1109/TCBB.2024.3306214
24. Zhou, M., Lin, W., & Tang, Y. (2023). Hybrid deep learning and genetic optimization for robust genomic cancer prediction. *Frontiers in Genetics, 14*, 1123871. https://doi.org/10.3389/fgene.2023.1123871