ISSN: 0067-2904

Classifying High-Dimensional Lung Cancer Data Using an Improved Deep Learning Model

Sarab M. Hameed *, Dhuha Abdulhadi Abduljabbar
Computer Science Department, College of Science, University of Baghdad, Baghdad, Iraq

Received:14/ 4/2025 Accepted: 5/ 10/2025 Published: xx

Abstract

Anomalous cell growth, known as cancer, can cause tumors, weaken the immune system, and cause fatal complications. Early detection of cancer makes treatment easier and contributes to lower mortality rates. Gene expression data contributes crucially to the classification of cancer in its early stages. Accurately classifying cancer is a complex problem because gene expression data is high-dimensional compared to the limited number of samples. This research proposes an improved Deep Learning model that integrates both Supervised Autoencoder (SAE) and Convolutional Neural Network (CNN), named as (SAE-CNN), to address this problem. Specifically, the SAE model was first developed to reduce the dimensionality of the dataset by identifying the most influential genes that will be considered as the main input data for the proposed effective CNN cancer classification model. The proposed cancer classification model is applied to several lung cancer gene expression microarray datasets that include distinct classes of cancer. The experimental results demonstrate that the proposed model outperforms existing state-of-the-art methods and proves its effectiveness in cancer classification, although the size of high-dimensional gene expression data is limited. Where Lung_Harvard2, Lung_Michigan, and Lung_Ontario datasets achieved a maximum accuracy of 100%, while Lung_Adenocarcinoma dataset achieved 90.48%. Moreover, the results demonstrate that the proposed model has the potential to help treat cancer and increase survival rates in the future.

Keywords: Lung cancer classification, High-dimensional gene expression data, Supervised Autoencoder, Convolutional Neural Network.

تصنيف بيانات سرطان الرئة عالية الأبعاد باستخدام نموذج التعلم العميق المحسّن

سراب مجيد حميد*، ضحى عبد الهادي عبد الجبار
قسم علوم الحاسبات، كلية العلوم، جامعة بغداد، بغداد، العراق

الخلاصة

يمكن أن يُسبب نمو الخلايا الشاذ، المعروف بالسرطان، أورامًا، ويُضعف جهاز المناعة، ويُسبب مضاعفات مميتة. يُسهّل الكشف المُبكر عن السرطان العلاج، ويُساهم في خفض معدلات الوفيات. تُساهم بيانات التعبير الجيني بشكل كبير في تصنيف السرطان في مراحله المُبكرة. يُعدّ التصنيف الدقيق لبيانات مرض السرطان مُشكلةً مُعقّدةً نظرًا لتعدد أبعاد بيانات التعبير الجيني مُقارنةً بالعدد المحدود من العينات. يقترح هذا البحث نموذجًا

*Email: sarab.m@sc.uobaghdad.edu.iq

مُحسّنًا للتعلم العميق يعتمد على دمج كلٍّ من النموذج المُشفّر التلقائي الخاضع للاشراف (SAE) ونموذج الشبكة العصبية التلافيفية (CNN) لمعالجة هذه المُشكلة. على وجه التحديد، طُوّر نموذج SAE في البداية لتقليل ابعاد مجموعة البيانات من خلال تحديد الجينات الأكثر تأثيرًا التي ستُعتبر بيانات الإدخال الرئيسية لنموذج تصنيف بيانات مرض السرطان المحسن المُقترح باستخدام نموذج CNN. يُطبّق نموذج تصنيف بيانات مرض السرطان المُقترح على العديد من مجموعات بيانات مصفوفة التعبير الجيني لسرطان الرئة، والتي تشمل أنواع مُختلفة من السرطان. تُظهر النتائج التجريبية أن نموذج المُقترح يتفوق على أحدث الطرق المُستخدمة حاليًا، ويُثبت فعاليته في تصنيف بيانات مرض السرطان، على الرغم من أن حجم بيانات التعبير الجيني عالية الأبعاد محدود.

1. Introduction

Cancer occurs when a cell experiences abnormal and uncontrolled growth due to changes at the cellular level. Cancer is one of the most dangerous diseases and one of the main causes of death in the world, according to a report issued by the World Prosperity Organization [1]. Lung cancer is one of the most common cancers in the world, causing the death of more than a million people every year. In addition to smoking, there are other risk factors such as exposure to second-hand smoke, occupational exposure to carcinogens, including radon gas, asbestos, air pollution, diesel exhaust, arsenic, and the role of family history in the development of the disease [2, 3]. Also, molecular and genetic studies have proven that there are various mutations and changes associated with the development of lung cancer [2]. Successful cancer diagnosis and treatment rely on accurate classification. Consequently, attention has increased due to advances in cancer classification [2, 4]. In general, many important biological applications, such as identifying disease-related genes, understanding biological cell functions, and predicting therapeutic uses, can be achieved by clustering and analyzing biological networks and protein networks [5], or by classifying gene expression data covering different classes of diseases, such as classifying high-dimensional lung cancer data. Cancer research has seen significant progress thanks to proteomic and genomic profiling techniques, which have enabled a deeper understanding of the molecular mechanisms responsible for carcinogenesis [6].

In the context of biomedicine, lung cancer classification relies on clinicopathological features. Gene expression microarrays have contributed to providing high throughput for determining genomic biomarkers used in cancer diagnosis and prediction [7, 8, 2]. Gene selection is challenging due to the properties of microarray data, which are structured, low sample size, high noise levels, and high dimensionality. It is essential to identify the best candidate genes for classification to solve this problem [9]. Deep learning (DL) models have a great importance in the medical field, focusing on addressing the cancer classification problem due to their ability to learn key features from raw data; hence, there is less need for manual feature engineering, which is especially useful when processing a wide range of tissue features. Developing a computer-assisted diagnosis system for medical applications is extremely challenging, and much research and funding have been granted for various diseases [**10**]. This paper aims to develop a model for lung cancer classification on high-dimensional gene expression data using deep learning techniques, namely Supervised Autoencoder (SAE) and Convolution Neural Network (CNN), coined as (SAE-CNN). To this end, the main contributions of this paper are:

- A new supervised autoencoder model-based gene expression reduction is proposed.
- Propose a CNN deep learning model to handle high-dimensional feature spaces efficiently, overcoming the exponential complexity related to dimensionality to classify lung cancer more accurately.

The structure of the paper is as follows: Section 2 presents some previous works on cancer classification based on determining informative genes. Section 3 provides the methods and

materials, including a discussion of the proposed DL models. Section 4 presents the results, facilitating a comprehensive discussion. Section 5 presents the conclusions and future work.

2. Related Work

Accurately classifying cancer types using microarray data is of paramount importance. Researchers have relied on a variety of AI techniques to analyze and examine gene expression data. The biggest challenge in cancer diagnosis is thought to be identifying genes that provide useful information. Therefore, selecting genes for study is a first step and a common practice in the cancer classification process [9]. This section presents some previous works on cancer classification-based determining informative genes, with a focus on the application of machine learning and deep learning models. Alsmadi [2] developed a modified version of AOA (Arithmetic Optimization Algorithm) that effectively worked on the high-dimensional dataset. In the proposed algorithm, the features are ranked according to their weights and employed to create a population, a set of individuals, for the AOA algorithm. Then, a local search algorithm is developed based on two different neighborhood approaches to enhance the AOA exploitation process. As a final point, the effectiveness of the developed methods was tested on lung cancer gene expression datasets. The efficiency of the proposed method in picking out the optimal subset of genes is demonstrated by its ability to keep the proportion of selected features between 10% and 25%, where the proposed method achieved a maximum accuracy of 100% for two out of five cancer datasets. In addition, this method contributes significantly to improving the accuracy of lung cancer prediction. However, the performance of the proposed algorithm is limited since the effectiveness of the initialization process depends on the ranking method.
Tabassum et al., 2024 [**11**] addressed the problem of accurate cancer classification using a dimensionality-reduction technique. First, the mutual information (MI) mechanism was used to select key vital genes. The ensemble learning model was then applied over the processed data that contained only the most influential genes (features) to develop an efficient model for the cancer classification problem. The bagging method was chosen as the ensemble technique in which the base classifiers are Multi-layer Perceptrons (MLPs). Experimental results showed that the suggested model outperformed existing methods, proving its effectiveness and efficiency in cancer classification despite the limited size of high-dimensional gene expression data, where it achieved 100% accuracy for three out of five cancer datasets. One limitation of this study is that the hidden layers do not adapt to the dimensions of the target sample data. Proper adjustment of the number of hidden layers can greatly improve the performance of the classifier.

Houssein et al., 2023 [**9**] offered a new classification approach combining the Support Vector Machine (SVM) algorithm with Runge Kutta optimizer to select genes of interest for the detection of cancer tissues. In order to process the high-dimensionality problem of microarray datasets, a pre-processing stage , the ReliefF method is applied. Experiments have shown that the RUN-SVM methodology outperforms other methods in achieving the highest accuracy rate. On the other hand, the developed method (RUN–SVM) presented a certain limitation in that it utilized an optimization technique based on randomization, which means that the features used may differ from one run to another. Thus, the set of features selected in one run may not be available in another run. Liu and Yao 2022 [**12**] proposed a method for gene selection based on KL variance, where some genes with the highest KL variance were selected as model features. Then they developed a deep neural network model with a loss function called the Focal Loss function. Meanwhile, they used the cross-validation method with a five-fold to select the best model and verify its effectiveness. The proposed KL divergence-based gene selection deep neural network model in this study has proven its efficiency and accuracy in lung cancer prediction, where the proposed model recorded an AUC of 0.99 across the validation

dataset. The authors noted that deep learning will gain increasing importance in disease diagnosis with great potential for development. However, models based on deep learning techniques require high computational costs; this problem can be addressed by using feature selection methods during the data pre-processing phase.

A hybrid model for gene selection was proposed by El Kafrawy et al., 2021 [**13**], termed SVM-mRMRe. Key genes were extracted from high-throughput microarrays by the proposed model, which combined filter-based, ensemble, and embedded methods by fusing embedded SVM coefficients (feature ranking) with ensemble mRMRe. The selected subset feature is evaluated through applying different types of classifiers, which are random forest (RF), KNN, NN, and SVM. As shown by experiments, the proposed method improved the current state of development in the field of classifying cancer tissue while relying on less data. In the same year, Qasem and Saeed [**14**] proposed a set of methods for gene selection and classification based on ranking and wrapper techniques. In the ranking methods, dimensionality was reduced using information gain measures up to 1% and 5%. In the wrapper methods, algorithms such as Naïve Bayes and K-nearest neighbors were applied with strategies such as Rank Search, Greedy Stepwise, and Best First. Several combinations have been studied since it is known that a single model cannot achieve the best results with different datasets and in all cases. Thus, integrating multiple feature selection methods with the use of diverse classification models can be more effective in making the final decision about the expected cancer types. However, there are no comprehensive evaluation criteria available for the dataset.

In general, classifying lung cancer data is a challenging task due to the data having the properties of microarray data, which are structured, described by low sample size (often less than a hundred), high dimensionality (often numbering in the thousands), and high noise levels. There is a need to identify the informative genes to solve the problem effectively with low computational costs, as well as to make physicians concentrate on a small set of genes that have physiological importance and are related to certain types of cancer, which makes it possible to conduct more cost-effective studies [**9**]. According to the reviewed studies, we can conclude that some strong and efficient gene reduction methods (i.e., Supervised AutoEncoder deep learning models) for lung cancer classification were overlooked in previous studies. The SAE-based feature reduction method can take into account capturing the underlying distribution of the sample and identifying anomalies that deviate significantly from that distribution by automatically encoding and decoding the data entered during the training phase. Therefore, this study proposes to develop a lung cancer classification model on high-dimensional gene expression data using deep learning techniques by integrating an efficient gene reduction method, namely Supervised Autoencoder (SAE) with Convolution Neural Network (CNN) as a powerful classification model.

3. Deep Learning Models

DL can be defined as a multi-layer neural network that learns from every sample it receives, grouping samples that share certain patterns by analyzing many parameters, and then accurately classifying the sample into its correct class based on the tested dataset. In this paper, two DL models, namely SAE and CNN, are adopted. In general, an autoencoder is a non-recurrent feed-forward neural network, much like a multilayer neural network. It is used for unsupervised learning to generate a concise and information-rich representation of data. It consists of one input layer, one output layer, and one or more hidden layers [**15, 16**]. In other words, an autoencoder receives the input data as a set of features and applies complex transformations to them, including compressing these features into a low-dimensional space. It then re-represents the original features using the compressed data. This can identify features with less influence

in the low-dimensional data (i.e., hidden units) as duplicates, which a group sparsity regularization can remove [17]. AE has been implemented and applied for several reasons, including its capability to extract valuable features by automatically encoding and decoding the input data during the training phase [18-21]. On the other hand, Supervised autoencoder (SAE) is an extension of the traditional autoencoder, where a supervised loss is added and the latent representation is taken as its input. In supervised learning, the goal is to learn a function capable of transforming input vectors into target vectors based on a limited set of independent, distributed data, with the goal of achieving high predictive power when dealing with new samples from the same distribution [22].

A convolutional neural network (CNN) is one of the most prominent networks. CNN has achieved remarkable success in many fields, such as natural language processing and computer vision, which has attracted great attention from academia and industry in recent years [23]. Based on the concept of visual perception, the CNN architecture is inspired [**24**]. In this architecture, a biological neuron corresponds to an artificial neuron, while the kernels of the CNN represent receptors that respond to different properties. Activation functions mimic the mechanism of neural signal transmission, where signals are transmitted only if they exceed a certain threshold. Optimizers and loss functions are invented to train the entire CNN system to learn the expected outcomes. Compared with traditional artificial neural networks, CNN has several advantages [23].

- Local connections: Each neuron connects to a limited number of neurons, but not all, in the previous layer, which helps reduce the number of parameters and increase the convergence speed.
- Weight sharing. It means having a set of connections that allow the same weights to be used and shared, further reducing the number of parameters.
- Reducing the sampling dimension: It refers to the pooling layer relying on the principle of local correlation of data to reduce the sample size of the data while retaining important information. It also helps in reducing the number of parameters by removing unimportant features.

4. The proposed Model-based Lung Cancer Classification

Given the labelled sample matrix $X = [x_1, \dots, x_m]^T \in R^{m \times d}$ where m is the number of labelled samples, and d is the dimensionality of the sample, i.e., the number of features. The main objective of this paper is to use a supervised autoencoder for supervised dimensionality reduction to select s $(s \leq d)$the most discriminative and informative features from X data. SAE is chosen to effectively reduce dimensionality, mitigate the impact of biological and technical noise, and extract features that are more discriminative for class labels, thereby addressing the challenges posed by the small sample sizes and high feature counts typical of gene expression datasets. CNN is utilized for lung cancer classification. The general block diagram of the proposed DL model is depicted in Figure 1. In the next sub-sections, we will provide details of the main components of the proposed model.

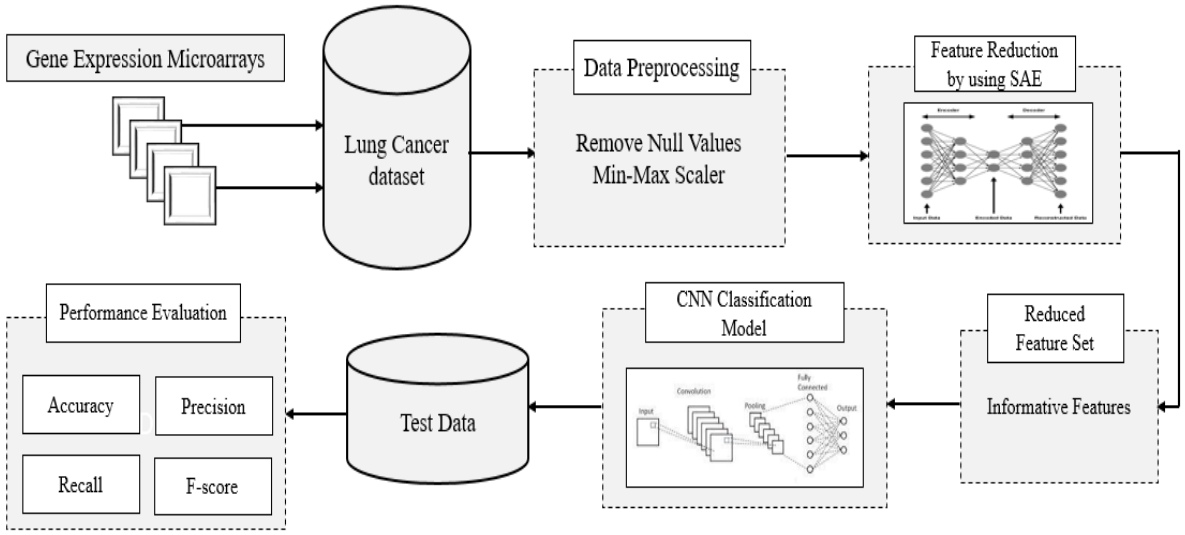

Figure 1: Block diagram of the proposed SAE-CNN model for lung cancer classification

4.1 Data Pre-processing

To ensure data is in the appropriate shape for training, testing, and evaluating deep learning models, two main pre-processing techniques are applied to facilitate data processing using SAE and CNN models. Firstly, it is essential to handle any missing values in the dataset properly to ensure that they do not negatively impact the performance of the model [**21**]. In these cases, the missing values are excluded. Secondly, the feature scaling method is applied to standardize the range of numerical features, preventing some features from controlling the learning process. This method also ensures the speed of convergence of the optimization algorithm and helps avoid numerical instability in the model. The standardization method of Min-Max scaling is used to linearly transform the numerical features. We refer to x_{min} as the minimum of the variable in the samples, while and x_{max} as the maximum. The Min-max algorithm determines the value of a variable, v, to the value, v', using the following equation: [**25**].

$$v' = \frac{v - x_{min}}{x_{max} - x_{min}} + x_{min} \quad (1)$$

After applying the Min-max algorithm, a variable in the dataset samples in the interval [x_{min}, x_{max}] has been scaled to $[0, 1]$ using a linear mapping.

4.2 The Proposed Supervised Autoencoder Model for Feature Reduction

In general, an autoencoder is a feedforward neural network that receives a set of features and outputs them after applying a series of transforms [**17**]. On the other hand, Supervised Autoencoder is an autoencoder model in which a supervised loss function is added to the model [22]. In a supervised learning environment, the aim is to learn a function that can transform input vectors $x \in R^m$ into target vectors $y \in R^d$. This function is trained on a limited set of independent, identically distributed data $(x_1, y_1), (x_2, y_2), ..., (x_t, y_t)$ to achieve high predictive power when encountering new samples from the same distribution [22]. In general, in a supervised autoencoder (SAE), the process of learning a compact representation of the data (encoding) is combined with its reconstruction (decoding) at the same time as the prediction or classification task is being performed, allowing the model to extract an efficient representation of the data and improve prediction accuracy. In this study, an SAE model is proposed for feature reduction to address the challenge of processing high-dimensional gene expression data by identifying meaningful and informative features for the tested microarray datasets. The following reasons demonstrate why an SAE was used for feature reduction in the proposed model:

1. Gene expression datasets typically involve numerous genes (features) for a limited number of samples, which makes traditional analysis methods difficult. Supervised Autoencoders compress the high-dimensional data into a lower-dimensional latent space, effectively reducing the number of features while preserving essential information.
2. Gene expression data often exhibits non-linear patterns and relationships between genes. Supervised Autoencoder is capable of modelling complex, non-linear relationships in the data by using non-linear activation functions.
3. Gene expression data is noisy due to biological variability, technical artifacts, and measurement errors. SAE can reconstruct the original data from a noisy input, helping to filter out irrelevant variations and retain key gene expression patterns.
4. The bottleneck layer of a supervised autoencoder captures the important biological features and discards redundant information, making it easier to interpret.
5. Gene expression datasets often combine data from multiple sources, which can present batch effects or other forms of heterogeneity. Supervised Autoencoder can handle this by learning representations that are less sensitive to such variations, improving the generalizability of the results.
6. Supervised Autoencoder is supervised learning, making it useful for gene expression datasets where labels are available for every sample.
The proposed SAE model consists of four densely connected layers for the encoder and decoder that produce a reconstruction. The encoder and decoder functions of the classical autoencoder are described in Equations 2 and 3, respectively.

$$f(X) = \sigma_1(XW^{(1)}) \quad (2)$$

$$X' = g\big(f(x)\big) = \sigma_2\big(f\big(X\big)W^{(2)}\big) \quad (3)$$

where σ_1, σ_2 are ReLU activation function of the hidden layer, and the activation function for the output layer is sigmoid. $\theta = \{W^{(1)}, W^{(2)}\}$ are weight parameters and $W_{ij}^{(l)}$ represents the connection parameter between neurons i in layer l and neuron j in layer $(l+1)$ [**17**]. In the learning process, the loss function, termed $Binary\ Cross-Entropy$, is minimized as formulated in Equation 4, which measures the difference between two probability distributions for a given random variable or set of events [**26**].

$$Binary\ Cross-Entropy = -(Xlog(g(f(X))) + (1-X)\ log\ log\ (1-g(f(X)))\ ) \quad (4)$$

When optimizing the above equation, we can obtain the observed matrix X compressed to the low-dimensional data $f(X)$ and yields the decoded matrix $X' = g\big(f(x)\big)$.
When using SAE, the classical autoencoder is extended by inserting supervision for the classification process [22]. In general, the SAE model consists of a multi-layer architecture, with encoding layers that reduce dimensionality and extract important features from the data, and decoding layers that reconstruct the data from the reduced representation. In addition, the model has two final outputs: one for data reconstruction and the other for binary classification. This architecture enables the SAE model to jointly learn an efficient representation of the data and achieve good classification accuracy.

4.3 The Proposed CNN Model for Lung Cancer Classification

The proposed architecture of the CNN model consists of three convolutional layers, two pooling layers, one flatten layer, three dropout layers, and one dense layer to process the high-dimensional lung cancer gene expression microarray datasets. The microarray gene expression data is converted from a two-dimensional (number of samples × number of features) to a three-dimensional (number of samples × number of features × 1) format by adding a single channel dimension in order to meet the requirements of Keras' one-dimensional (Conv1D) convolutional network layers, which require a three-dimensional input**.** Thereafter, in the convolution layers, a set of weights is employed called kernels of size 5 for the first two layers

and 3 for the last layer, where they are multiplied by inputs to produce a new feature map when the activation function is set to ReLU for the first layer and tanh for the remaining layers, with kernel_regularizer set to ℓ_2 (0.001) for all ConvID layers plus the first dense layer. Then, a pooling layer is applied to reduce the number of trainable parameters. A flattening layer is used to transform the 3D vector into 1D inputs for the dense layers. Algorithm 1 describes the steps of the proposed DL model based on combining a supervised autoencoder and a convolutional neural network.

In general, a supervised autoencoder (SAE) consists of a multi-layer architecture that begins with an input layer that processes raw data. This is followed by encoder layers that reduce dimensionality to extract important features and a reduced representation of the data. Decoder layers then reconstruct the original data from the reduced representation, aiming to preserve key information. Additionally, the model contains a classification layer that enables supervised learning tasks, such as classifying data into two classes. This architecture enables the model to learn an efficient representation of the data while improving classification performance, making it suitable for use in dimensionality reduction tasks followed by accurate classification. Convolutional Neural Network (CNN) includes Convolution Layers to extract the basic features, Pooling Layers to reduce the dimensionality, kernel regularizer to add ℓ_2 regularization to the weights to reduce overfitting by decreasing the weight values during training, and finally Fully Connected Layers to transform the features into classification results.

Algorithm 1: The proposed SAE-CNN model for lung cancer classification	
Input: ● Lung cancer gene expression microarray datasets: $X = [x_1, \dots, x_m]^T$ ● Define the hyperparameters. **Output:** ● DL model for lung cancer classification.	
1:	Handle missing values in X by removing null values.
2:	Scale Lung cancer gene expression microarray datasets using Eq. (1)
3:	Build and compile SAE model with the defined hyperparameters as inputs. Use the developed SAE model to reduce the dimensionality of the data and output X'.
4:	Split the pre-processed dataset, X' into train set X'_{tr}, and test set X'_{ts} using 4-fold cross-validation strategy.
5:	Build and compile CNN model with the defined hyperparameters as input.
6:	Train the model on the training lung cancer data, X'_{tr} using the determined hyperparameters.
7:	Evaluate the trained model on the test lung cancer data, X'_{ts} in terms of precision, recall, F-measure, and accuracy.
8:	print the mean of the defined evaluation metrics.

4.4 Model architecture and training procedures

In this section, we review the model settings and hyperparameters used in the experiments, including the architectural structure and training options such as activation functions, optimizers, batch size, number of epochs, and learning rate. The values of hyperparameters of SAE-CNN are based on their common settings noted in the earlier studies. The hyperparameters of the SAE model included:

1. Activation function: ReLU function with the encoding and decoding layers, and sigmoid with the output layers for reconstruction and binary classification.
2. The number of neurons per the first and second encode and decode dense layers was determined based on the maximum and minimum number of features reported in the reference

[**2**], as follows: the minimum number of features is: **1318, 911, 1323, 615** while the maximum number of features is: **1925, 1364, 2975, 1158** for the Lung_Harvard2, Lung_Michigan, Lung_Adenocarcinoma, and Lung_Ontario, respectively.

3. SAE has two final outputs, each of which is an independent Dense layer: the first is dedicated to data reconstruction, and the second is dedicated to binary classification.
4. Optimization Function: RMSprop
5. Batch Size: 32
6. Epochs: 50 with $patience = 5$ for the EarlyStopping function.
7. Loss Function: Binary cross-entropy, where two main loss functions are used: one to reconstruct the original data, and the other to evaluate the accuracy of the binary classification, with each given equal weight during training.
8. Learning Rate: 0.0005.

The hyperparameters of the CNN model included:

1. Activation functions per the three convolution layers are ReLU, tanh, and tanh, respectively, with a kernel regularizer.
2. The kernel size per the three convolution layers is $5, 5$ with $filters = 32$, and 3 with $filters = 64$, respectively.
3. Dropout layers are $0.3, 0.4$, and 0.5, respectively.
4. Pool size per the two MaxPooling1D layers is 2.
5. Dense layer with 128 neurons, ReLU activation function, and kernel regularizer; while the second dense layer with a single neuron and sigmoid activation function.
6. Optimization Function: RMSprop with $learning\_rate = 0.0005$, $rho = 0.9$, $epsilon = 1e - 08$.
7. Batch Size: 50, epochs: 500 with a class weight factor.
8. Loss Function: binary cross-entropy.

5. Experimental Results

5.1 Benchmark datasets

In this paper, four public lung cancer gene expression microarray datasets were selected to verify the performance of the proposed DL model. Table 1 shows the general characteristics of the datasets. The details of each dataset are as follows:

1. Lung_Harvard2: MPM (malignant pleural mesothelioma) and ADCA (adenocarcinoma of the lung) were divided into several classes. The Lung_Harvard2 dataset contains 181 tissue samples, of which 150 are ADCA samples and 31 are MPM samples. 12,533 genes were used to describe each sample [**27**].
2. Lung_Michigan: Ten non-neoplastic lung samples and 86 primary lung adenocarcinoma samples are also included. Each sample is characterized by 7129 genes [28].
3. Lung_Adenocarcinoma: This dataset includes 7,129 previously processed genes from 86 primary lung adenocarcinoma patients, of whom 62 are still alive and 24 have died [**7**, 29].
4. Lung_Ontario: The data relates to the gene expression of 39 different NSCLC samples. Twenty-four patients in these samples exhibited distant metastases (classified as “relapse”) or local tumor relapses. The remaining 15 patients, according to clinical examinations and radiological findings, were free of disease (classified as “non-relapse”). 2880 genes were used to term the processed data [30].

To standardize the data format and facilitate processing, pre-processing was performed that included encoding all data outputs into binary numeric values ($0\ and\ 1$), where each category was represented by one of these two values, ensuring its efficient use in classification and analysis processes.

Table.1: General characteristics of lung cancer datasets

Dataset Name	# of Samples	# of Features	Class Distribution
Lung_Harvard2	181	12533	150- ADCA 31-MPM
Lung_Michigan	96	7129	72- Adenocarcinomas 24- non-neoplastic
Lung_Adenocarcinoma	86	7129	62- Alive 24- Died
Lung_Ontario	39	2880	24- relapse 15- non-relapse

5.2 Evaluation Metrics

The proposed SAE-CNN model was compared to the other methods using four different measures: accuracy, precision, recall, and F-measure. The correctness of classification is evaluated by counting the number of samples that are correctly classified as being in the true class (true positives), the number of samples that are correctly identified as being out of the class (true negatives), the number of samples that are incorrectly classified as belonging to the class (false positives), and the number of samples that are ignored despite being in the class (false negatives). The values of these four terms form what is known as the confusion matrix [**31**].

Table 2 presents the most often used measures for binary classification based on the values of the confusion matrix. In the case of every dataset, the results obtained are shown with 4-fold cross-validation for robustness.

Table 2: Summary of evaluation metrics used in this paper

Measure	Formula	Interpretation
Accuracy (Acc)	$\frac{tp + tn}{tp + fn + fp + tn}$	It is defined as the ratio of correctly classified samples to the total samples in the dataset.
Precision (P)	$\frac{tp}{tp + fp}$	The fraction of the true-positive classifications out of all the positive classifications
Recall (R)	$\frac{tp}{tp + fn}$	The fraction of the true-positive classifications out of all the true classifications.
F-measure (F1)	$2 \times \frac{precision \times Recall}{precision + Recall}$	It reflects the trade-off between precision and recall and measures the model's performance by taking into account both recall and precision.

5.3 Results Discussion

In this paper, we found that the combination of supervised autoencoder and convolutional neural network models proposed has achieved better results than the counterpart classification models in lung cancer classification. Accurate and early detection of lung cancer is of great importance because of the high incidence and mortality rates of this type of cancer. Therefore, computer AI technologies play a significant role in enhancing the prediction accuracy of lung cancer disease [**12**]. All experiments were performed on a Windows 10 Pro (64-bit Operating system) with 16.0 GB RAM and 250 GB SSD. Python version: 3.11.7 was used on an Anaconda version conda 24.11.2 for implementing the proposed model.

Table 3 reports the comparative performance of the Supervised Autoencoder (SAE), standard Autoencoder (AE), and full feature set across the four lung cancer datasets when the CNN model was applied. The results stated in Table 3 reveal several remarkable points.

Firstly, it was noted that when using AE-based or SAE-based feature reduction methods, the computation time is highly improved since the useless, redundant, and abnormal features are excluded.

Secondly, the results demonstrate that SAE with CNN consistently outperforms AE with CNN and even outperforms the full feature set, particularly for Lung_Harvard2, Lung_Adenocarcinoma, and Lung_Ontario; while on the Lung_Michigan dataset, the model achieved optimal performance in all test cases. This is due to SAE integrating label information into the learning process by optimizing both reconstruction loss and classification loss, enabling the latent space to preserve discriminative features critical for accurate classification. Whereas AE tended to compress CNN features in ways that diminished their discriminative power. These results demonstrate that SAE is better suited for dimensionality reduction in classification tasks, especially in complex biomedical datasets where subtle inter-class differences must be preserved.

Thirdly, one can observe that the dimensionality reduction with SAE is highly valuable in complex datasets, namely Lung_Harvard2**,** Lung_Adenocarcinoma, and Lung_Ontario, in terms of speed and accuracy, where irrelevant or noisy features may be impacting classification performance when using the whole feature set. While in other datasets, the model performs perfectly regardless of the number of features.

Table 3: Results of the proposed CNN model across four lung cancer datasets with whole features and after applying AE and SAE as feature reduction models

Dataset	Model	# of Feature	% of Feature	% Time improvement	Mean P	Mean R	Mean F1	Mean Acc	Min Acc	Max Acc	Std Acc
Lung_Harvard2	CNN	12533	100%	_	99.29	100	99.63	99.44	97.78	100	0.0096
	AE-CNN	1318	10.516 %	92.12%	98.57	98.65	98.57	97.78	95.56	100	0.0222
	SAE-CNN	1318	10.516 %	92.47%	**100**	**100**	**100**	**100**	**100**	**100**	**0.00**
Lung_Michigan	CNN	7129	100%	_	**100**	**100**	**100**	**100**	**100**	**100**	**0.00**
	AE-CNN	911	12.778 %	86.56%	**100**	**100**	**100**	**100**	**100**	**100**	**0.00**
	SAE-CNN	911	12.778 %	86.54%	**100**	**100**	**100**	**100**	**100**	**100**	**0.00**
Lung_Adenoca-rcinoma	CNN	7129	100%	_	73.69	**90.31**	81.09	69.86	59.09	76.19	0.0646
	AE-CNN	1323	18.558 %	80.03%	72.80	77.19	74.49	62.88	54.55	76.19	0.0837
	SAE-CNN	1323	18.558 %	80.15%	**88.67**	85.31	**86.53**	**81.44**	**76.19**	**90.48**	**0.0563**
Lung_Ontario	CNN	2880	100%	_	76.25	**85.42**	78.60	79.17	66.67	**100**	0.1299
	AE-CNN	615	21.354 %	11.95%	48.13	60.42	50.57	68.61	44.44	80	0.1454
	SAE-CNN	615	21.354 %	11.73%	**78.75**	**85.42**	**80.68**	**81.94**	**70**	**100**	**0.1107**

Table 4 shows the performance comparison of the proposed CNN model with full feature set in terms of precision, recall, F1-score, and accuracy on different sizes of datasets namely lung_adenocarcinoma, lung_michigan, lung_ontario, and lung_harvard2 with the baseline classifiers: KNN (K-Nearest Neighbor), SVM (Support Vector Machine), DT (Decision Trees), RF (Random Forest), and NB (Naive Bayes). The proposed model provides competitive or superior performance compared with baseline classifiers across the four lung cancer datasets. With a larger sample size, namely Lung_Harvard2, which has very high-dimensional features,

the performance is high and consistent across all models; this dataset appears to be well-suited for both the traditional and the proposed model. The proposed model and RF achieved the best performance, as it was able to correctly detect and classify most samples. On the other hand, NB and SVM performed a notable performance, followed by DT and KNN, respectively. In the case of Lung Michigan data, the proposed model outperforms other models by reaching 100% in all metrics. This may be due to the dataset being less noisy or having well-separated classes. The rest of the classifiers achieved good and comparable results, except the DT model, which returned the lowest performance in terms of F1-score.

For Lung_Adenocarcinoma, which contains 86 samples, the performance of the proposed CNN drops significantly, indicating greater class overlap and misaligned features. The results of KNN, SVM, and RF show moderate success, but do not reach the high accuracy of Lung_Harvard2 and Lung Michigan datasets. This is because the dataset has a complex feature space that deep learning models handle better. The results in Table 4 indicated that all the classifier models were unable to classify all samples into their correct categories, where even the proposed model only reached a moderate accuracy (69.85) and a comparable $F_1 - score$ (81.09), while the NB classifier showed moderate performance in terms of F1-score and accuracy due to the feature dimension being too high, and the NB cannot be fully learned, resulting in low classification accuracy. On the other hand, DT recorded an accuracy equal to 69.97 due to DT frequently overfitting because of insufficient data and extraneous characteristics, resulting in intricate decision boundaries and inadequate generalization. For a Lung_Ontario dataset, a smaller dataset size, the F1-score for all counterpart models is low, which means that the dataset may have high noise. The proposed model reaches the highest results in terms of precision (76.25), recall (85.42), F1-score (78.60), and accuracy (79.17) among all the counterpart models. This indicates that while traditional machine learning methods such as KNN, RF, and NB perform well in some cases, the proposed model tends to generalize better, especially on larger datasets.

Table 4: Results of the proposed model and other ML methods without feature reduction (full set of features)

Dataset	Model	Precision	Recall	F1-score	Accuracy
Lung_Harvard2	KNN	90.25	**100**	94.85	91.15
	SVM	98.65	98.70	98.66	97.79
	DT	94.88	96.67	95.58	92.81
	RF	99.26	**100**	**99.63**	**99.44**
	NB	98.65	**100**	99.32	98.89
	The proposed CNN model	**99.29**	**100**	**99.63**	**99.44**
Lung_Michigan	KNN	87.50	93.75	88.10	97.92
	SVM	100	87.50	91.67	97.92
	DT	79.17	79.17	75.95	93.75
	RF	**100**	79.17	86.67	96.88
	NB	**100**	79.17	86.67	96.88
	The proposed CNN model	**100**	**100**	**100**	**100**
Lung_Adenoca-rcinoma	KNN	76.94	91.77	83.62	**74.40**
	SVM	72.13	**100**	**83.78**	72.13
	DT	**79.54**	80.52	79.57	69.97
	RF	72.13	**100**	**83.78**	72.13
	NB	75.82	80.83	78.14	67.48
	The proposed CNN model	73.69	90.31	81.09	69.86
Lung_Ontario	KNN	71.88	64.58	60.06	68.61
	SVM	34.38	31.25	23.64	58.61
	DT	55.71	54.17	49.85	71.39
	RF	60.71	45.83	44.17	68.89
	NB	51.96	75	56.79	61.39
	The proposed CNN model	**76.25**	**85.42**	**78.60**	**79.17**

Next, the proposed CNN with SAE model was evaluated with seven optimization algorithms [**2**], namely MAOA ("Modified Arithmetic Optimization Algorithm"), AOA ("Arithmetic Optimization Algorithm"), PSO ("Particle Swarm Optimization"), GWO, MFO ("Moth-Flame

Optimization"), FA ("Firefly Algorithm") and WOA ("Whale optimization algorithm"). It is worth noting that the results of all counterpart optimization algorithms were taken from the reference [2].

Table 5 presents the comparison results between the proposed model and seven counterpart optimization algorithms in terms of the minimum (min), maximum (max), average ($mean$), and standard deviation (std) of the accuracy metric. As shown in Table 5, the proposed model provides competitive or superior performance compared with the optimization algorithms. In general, the proposed model surpassed all other algorithms in terms of max accuracy across all four test datasets. On the other hand, most algorithms achieved high performance for Harvard2 and Lung_Michigan datasets. On the Lung_Harvard2 dataset, the proposed SAE-CNN model, PSO, and GWO algorithms outperformed all other algorithms in terms of $min, mean\ and\ std$ accuracy. As for Lung_Michigan dataset, the proposed model and all the optimization algorithms, except FFA, showed optimal performance in all accuracy scores. Lastly, a comparable performance was achieved by the proposed model for Lung_Adenocarcinoma and the Ontario dataset, scoring the third highest results compared to other algorithms in terms of $mean$ accuracy.

Table 5: Comparing the performance of the proposed model with seven counterpart optimization algorithms [2] in terms of Min, Max, Mean, and standard deviation (Std) accuracy

Datasets	Accuracy	MAOA	PSO	GWO	MFO	WOA	FFA	AOA	SAE-CNN
Lung_Harvard 2	Min	97.22	**100**	**100**	97.22	97.22	**100**	97.22	**100**
	Max	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
	Mean	99.07	**100**	**100**	99.07	99.07	**100**	99.07	**100**
	Std	0.02	**0.00**	**0.00**	0.02	0.02	**0.00**	0.02	**0.00**
Lung_Michigan	Min	**100**	**100**	**100**	**100**	**100**	94.74	**100**	**100**
	Max	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
	Mean	**100**	**100**	**100**	**100**	**100**	98.25	**100**	**100**
	Std	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.03	**0.00**	**0.00**
Lung_Adenoca -rcinoma	Min	**76.47**	70.59	**76.47**	70.59	70.59	58.82	70.59	76.19
	Max	88.24	88.24	88.24	88.24	88.24	82.35	88.24	**90.48**
	Mean	**84.12**	77.65	80.59	82.94	79.41	75.88	78.24	81.44
	Std	0.05	0.05	**0.04**	0.06	0.06	0.08	0.06	0.0563
Lung_Ontario	Min	75	75	75	75	75	75	75	70
	Max	87.50	75	87.50	87.50	87.50	87.50	87.50	**100**
	Mean	81.25	75	81.25	76.25	**85**	81.25	83.92	81.94
	Std	0.07	**0.00**	0.07	0.04	0.05	0.07	0.04	0.1107

6. Conclusions

Profiling of gene expression for early cancer diagnosis is a recent approach anticipated to enhance the early detection and management of diverse types of cancer. In this paper, an improved deep-learning model based on integrating both SAE and CNN is proposed for an effective cancer classifier. The effectiveness of the proposed deep learning model on four diverse cancer datasets is evaluated and compared with the baseline machine learning and optimization models. The proposed model achieved promising accuracy results compared to other corresponding methods on all datasets that have different characteristics regarding the number of features and samples. In the future, the model can be combined with traditional

classifiers to enhance its robustness with small and noisy datasets. Moreover, statistical tests can be used to assess the biological significance of selected genes compared to biomarkers. Additionally, the performance of the proposed SAE-CNN can be enhanced by adopting transformer-based models.

References

[1] V. Kumar, C. Prabha, P. Sharma, N. Mittal, S. S. Askar, and M. Abouhawwash, "Unified deep learning models for enhanced lung cancer prediction with ResNet-50–101 and EfficientNet-B3 using DICOM images," *BMC Medical Imaging*, vol. 24, no. 1, p. 63, 2024.

[2] M. K. Alsmadi, "Hybrid Gene Selection Methods for High-Dimensional Lung Cancer Data Using Improved Arithmetic Optimization Algorithm. Computers," *Materials & Continua,* vol. 79, no. 3, pp. 5175-5200, 2024.

[3] K. MS, H. Rajaguru, and A. R. Nair, "Enhancement of Classifier Performance with Adam and RanAdam Hyper-Parameter Tuning for Lung Cancer Detection from Microarray Data—In Pursuit of Precision," *Bioengineering*, vol. 11, no. 4, p. 314, 2024.

[4] P. Guan, D. Huang, M. He, and B. Zhou, "Lung cancer gene expression database analysis incorporating prior knowledge with support vector machine-based classification method," *Journal of experimental & clinical cancer research*, vol. 28, pp. 1-7, 2009.

[5] D. A. Abduljabbar, "Overlapping complexes detection within protein interaction networks using improved genetic algorithm," *Human Gene*, vol. 40, p. 201291, 2024.

[6] S. N. Tambe-Jagtap and Z. A. Jaaz, "Integrative genomic and proteomic profiling for personalized oncological treatments: enhancing therapeutic efficacy and reducing adverse effects in breast cancer patients". *SHIFAA*, vol. *2023*, pp. 1-9.

[7] D. G. Beer, S. L. Kardia, C. C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, ... and S. Hanash, "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature medicine*, vol. 8, no. 8, pp. 816-824, 2002. doi: 10.1038/nm733.

[8] S. Ramaswamy, K. N. Ross, E. S. Lander, and T. R. Golub, "A molecular signature of metastasis in primary solid tumors," *Nature genetics*, vol. 33, no. 1, pp. 49-54, 2003. doi: 10.1038/ng1060.

[9] E. H. Houssein, H. N. Hassan, N. A. Samee, and M. M. Jamjoom, "A Novel Hybrid Runge Kutta Optimizer with Support Vector Machine on Gene Expression Data for Cancer Classification," *Diagnostics*, vol. 13, no. 9, p. 1621, 2023.

[10] M. A. Thanoon, M.A. Zulkifley, M.A.A. Mohd Zainuri, and S.R. Abdani, "A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images," *Diagnostics*, vol. 13, p. 2617, 2023. https://doi.org/10.3390/ iagnostics13162617

[11] N. Tabassum, M. A. S. Kamal, M. A. H. Akhand, and K. Yamada, "Cancer Classification from Gene Expression Using Ensemble Learning with an Influential Feature Selection Technique," *BioMedInformatics*, vol. 4, no. 2, pp. 1275-1288, 2024.

[12] S. Liu, and W. Yao, "Prediction of Lung Cancer Using Gene Expression and Deep Learning with KL Divergence Gene Selection," *BMC bioinformatics*, vol. 23, no. 1, p. 175, 2022.

[13] P. El Kafrawy, H. Fathi, M. Qaraad, A. K. Kelany, and X. Chen, "An efficient SVM-based feature selection model for cancer classification using high-dimensional microarray data," *IEEE Access*, vol. 9, pp. 155353-155369, 2021.

[14] S. N. Qasem and F. Saeed, "Hybrid Feature Selection and Ensemble Learning Methods for Gene Selection and Cancer Classification," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, 2021.

[15] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, "Extracting and Composing Robust Features with Denoising Autoencoders," In *Proceedings of the 25th international conference on Machine learning*, pp. 1096-1103, 2008.

[16] D. Tomar, Y. Prasad, M. K. Thakur, and K. K. Biswas, "Feature Selection Using Autoencoders," In 2017 I*nternational Conference on Machine Learning and Data Science (MLDS)*, pp. 56-60, 2017. IEEE.

[17] K. Han, Y. Wang, C. Zhang, C. Li, and C. Xu, "Autoencoder Inspired Unsupervised Feature Selection," In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 2941-2945, 2018. IEEE.

[18] R. Ball, H. Kruger, and L. Drevin, "Anomaly Detection Using Autoencoders With Network Analysis Features," *ORiON*, vol. 39, no. 1, 2023.

[19] W. Hilal, S. A. Gadsden, and J. Yawney, "Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances," *Expert systems with applications*, vol. 193, p. 116429, 2022.

[20] M. S. Kwon, Y. G. Moon, B. Lee, and J. H. Noh, "Autoencoders with Exponential Deviation Loss for Weakly Supervised Anomaly Detection," *Pattern Recognition Letters*, vol. 171, pp. 131-137, 2023.

[21] S. S. Sulaiman, I. Nadher, and S. M. Hameed, "Credit Card Fraud Detection Using an Autoencoder Model with New Loss Function," *International Journal of Intelligent Engineering and Systems*, vol. 17, no. 5, pp. 210-220, 2024. DOI: 10.22266/ijies2024.1031.18

[22] Le, L., Patterson, A., and White, M. Supervised autoencoders: improving generalization performance with unsupervised regularizers. In *Proceedings of the 32nd International Conference on Neural Information Processing System*s 107–117 (Curran Associates Inc., Red Hook, NY, USA, 2018).

[23] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 12, pp. 6999-7019, 2021.

[24] D. H. Hubel and T. N. Wiesel, "Receptive Fields, Binocular Interaction and Functional Architecture In The Cat's Visual Cortex," *The Journal of Physiology*, vol. 160, no. 1, p.106, 1962.

[25] X. H. Cao, I. Stojkovic, and Z. Obradovic, "A Robust Data Scaling Algorithm to Improve Classification Accuracies in Biomedical Data," *BMC Bioinformatics*, vol. 17, pp. 1-10, 2016.

[26] S. Jadon, "A Survey of Loss Functions for Semantic Segmentation," In *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*, pp. 1-7, 2020. IEEE.

[27] G. J. Gordon, R. V. Jensen, L. L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, ... and R. Bueno, "Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma," *Cancer research*, vol. 62, no. 17, pp. 4963-4967, 2002.

[28] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, and M. Meyerson, "Classification of Human Lung Carcinomas by mRNA Expression Profiling Reveals Distinct Adenocarcinoma Subclasses," *Proceedings of the National Academy of Sciences*, vol. 98, no. 24, pp. 13790-13795, 2001. doi: 10.1073/pnas. 191502998.

[29] P. Shi, S. Ray, Q. Zhu, and M. A. Kon, "Top Scoring Pairs for Feature Selection in Machine Learning and Applications to Cancer Outcome Prediction," *BMC bioinformatics*, vol. 12, no. 1, pp. 1-15, 2011. doi: 10.1186/1471-2105-12-375.

[30] D. A. Wigle, I. Jurisica, N. Radulovich, M. Pintilie, J. Rossant, N. Liu, and M. S. Tsao, "Molecular Profiling of Non-Small Cell Lung Cancer and Correlation with Disease-Free Survival," *Cancer Research*, vol. 62, no. 11, pp. 3005-3008, 2002.

[31] M. Sokolova and G. Lapalme, "A Systematic Analysis of Performance Measures for Classification Tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427-437, 2009.