ISSN: 2093-5374 / E-ISSN: 2093-5382

Enhanced Accuracy for Lung Adenocarcinoma (LUAD) Prediction based UMAP Feature Using Artificial Neural Network

B. Jyothi[1*], and Dr.L. Mary Gladence[2]

[1*]Research Scholar, Department of CSE, Sathyabama Institute of Science and Technology, Chennai, India. jyothiniraj28@gmail.com, https://orcid.org/0009-0006-2578-5715

[2]Professor, School of Computing, Sathyabama Institute of Science and Technology, Chennai, India. marygladence.it@Sathyabama.ac.in, https://orcid.org/0000-0002-6767-6537

Received: August 06, 2024; Revised: September 17, 2024; Accepted: October 14, 2024; Published: December 30, 2024

Abstract

The Lung adenocarcinoma (LUAD), a common histopathological manifestation of carcinoma of the lungs as well as a variant, grouped as Non-Small Cell Lung Cancer (NSCLC), makes up 45-5% of every instance of cancer of the lungs. Different variables, notably environmental exposures and genetic makeup have been discovered to contribute to the onset and advancement of LUAD. It has been shown that combining the expression of genes with other information can help detect lung cancer patients. It offers several more perspectives that improve the categorization of cancers. Based on the results, it is believed that identifying the genes that have an extensive expression in malignant cells as opposed to typical ones is a difficulty, which calls for the application of computational tools. Applied computing techniques encountered more problems with this data set because of its elevated resolution and small sample size. Many supervised and unsupervised educational strategies have been developed for the classification of GED cancer. Since the most important traits remain unidentified, ML techniques have not been very successful in reducing dimensionality or classifying malignancy in GED patients. This research aims to address these challenges by leveraging the power of Uniform Manifold Approximation and Projection (UMAP) to enhance the feature space representation. This paper offers a unique Artificial Neural Network (ANN) model to predict, particularly LUAD types of cancer among the other gene expression data such as BRCA, COAD, KIRC, and PRAD cancer. The results of the proposed UMAP with ANN model 3 demonstrate for the detection of LUAD data when evaluated using performance measures has the highest accuracy of 99.53%.

Keywords: LUAD, Prediction, UMAP, ANN, Gene Expression, Cancer.

1 Introduction

Cancer is known to be one of the largest threats to humanity (Filipp, 2017; Archer et al., 2016). Data from epidemiology have shown that cancer is one of the leading fatalities in human beings, next only to transmittable, coronary artery disease, and neurological illnesses; this presents a significant risk to the wellness of humans (Vos et al., 2016). The expression of genes serves as one of the most widely used

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 15, number: 4 (December), pp. 380-394. DOI: *10.58346/JOWUA.2024.I4.026*

*Corresponding author: Research scholar, Department of CSE, Sathyabama Institute of Science and Technology, Chennai, India.

factors in categorizing cancer. Gene expression, or the transcriptome reflecting the actively transcribed genes at any point, can be used to ascertain the physiological condition and genetic function within biological structures (Liu et al., 2018; Saritha & Gunasundari, 2024).

According to statistics, carcinoma of the lungs is the next most prevalent disease in both individuals to be diagnosed and remains the most common cancer that results in mortality (Yuan et al., 2020; Huang et al., 2023). The Small Cell Carcinoma of The Lungs (SCLC), grows quickly but is uncommon, and NSCLC which grows slowly and is the most prevalent type of lung cancer identified, are the two primary subtypes of the disease. Chest pain, breathing difficulties, weight loss, and a chronic cough are signs of lung cancer. Physical examinations, several imaging modalities, and molecular screening to identify precise genetic alterations or indicators to aid in the selection of the most effective treatment options are among the diagnostic approaches for lung cancer (World Health Organization, 2023; Bodor et al., 2020; Ginn et al., 2020). Advances in genome profiling methods in recent decades have made considerable progress in our understanding of the genetic basis of cancer formation, which significantly improved the comprehension of cancer development at the molecular level and helped in cancer treatments including lung cancer by identification of biomarkers (Cai et al., 2019; Ramakrishnan et al., 2019; Alzaidi, 2024).

Gene expression profiling has provided valuable information about gene activities and describes the current physiology of the cell. It has been successfully applied to aid in the prognosis and preliminary diagnosis of various kinds of cancer (Osama et al., 2023; Alharbi & Vakanski, 2023). A useful method for measuring gene expression is single-cell RNA sequencing (scRNA-Seq), a potent technique in molecular science for analyzing gene transcription at the single-cell level (Xue et al., 2023). The process of translating the genetic codes found in DNA into proteins and other molecules is known as gene expression. DNA transcription is a fundamental process in molecular biology where information from a DNA sequence is transcribed into RNA. This process is a key step in gene expression, during which the genetic instructions stored in DNA are used to synthesize ribonucleic acid RNA molecules (Mathew & Asha, 2024).

A few more details that improve cancer categorization are provided by gene expression data (GED) (Thakur et al., 2021). Based on the findings, it is believed that identifying genes that are highly expressed in malignant cells as opposed to normal ones is a problem that calls for the application of computerized techniques. These data provided further challenges for the use of computational techniques because of their high resolution and small sample size. Analyzing the gene expression data is still challenging due to its many characteristics such as high dimensionality, complexity, and heterogeneity. Several unsupervised as well as supervised algorithms have been developed for the GED-based cancer classification. Deep Learning (DL) technologies are overcoming the limitations of traditional Machine Learning (ML) techniques while analyzing the data for cancer (Khalsan et al., 2022). Since the most important traits remain unidentified, ML techniques have not been very successful in reducing dimensionality or classifying malignancy in GED patients. Conversely, DL techniques have demonstrated superior performance in these two domains. GED offers additional attributes and fewer samples. As a result, ML models generate multiple repeated learners that consistently produce low-quality results regardless of the input that is provided to the computer.

It is believed that DL is a variant of ML. Due to its numerous advantages, DL strategies yield better results than models created with ML, requisite information and materials are accessible, and novel approaches are always being developed. DL is a rapidly expanding discipline in the modern day. The foundation of DL models is Stochastic Gradient Descent (SGD). Even while the ML models produce excellent results, they lag behind in many domains, such as object and speech detection. This serves as one of the causes for the DL field. ANN, which is impacted by the brain's organic neurons, gave rise to

DL models. The following three layers make up an ANN: an input layer, a layer that is hidden, and an output layer containing several nodes. Although generality is absent in monitored operations, these models acquire knowledge by means of a recurrent learning procedure called backpropagation. Similar to our own networks of neurons, ANNs provide a good method for resolving predictions and classification problems. An ANN is a computer framework modelled after the design and workings of genuine neural networks. In addition to the input and output layers, neural systems typically have hidden layers that convert information into variables that the resulting layer may utilize. The ANN model, a neural network used to identify cancer, goes through two stages: validation and training. A collection of data is used to train the network initially. The weights of the links between synapses are then fixed, allowing the structure of the network to be verified to provide classifiers for an entirely new set of data.

In this study, we focused on LUAD, which is the primary type of lung cancer diagnosis, and introduced a novel ANN model. This paper makes the following essential contributions.

1. Collection of datasets from open-source data repositories.
2. Performing data cleansing and pre-processing methods to render error-free data for modeling.
3. Performing dimensionality reduction on the data using the UMAP algorithm tuned by enhancing its parameters.
4. Build and compare LC-based novel ANN models on the reduced feature space rendered from tuned UMAP algorithm, using various Python libraries.
5. Conduct a comparative analysis for three ANN models that differ based on the neuron count in each layer to assess the effectiveness of tuned UMAP for the classification of lung cancer genes.

This research article is structured into three sections: Section 1 provides a broad review of the issue; and Section 2 describes the study's methodologies. Section 3 goes into detail about the experiments performed and the results obtained. Section 4 summarizes the experimental techniques, findings, and recommendations for future research, and serves as the paper's conclusion.

2 Literature Review

Mohamed & Ezugwu, (2024) describe proposing an improved DL model. This model presented an improved DL model for diagnosing the different stages in lung cancer (I, II, III, and IV) utilizing, a precisely selected multi-omics dataset. Nonetheless, the large complexity of the information and the inclusion of unbalanced class features made, building the prediction model with integrated heterogeneous datasets difficult. As a result, this paper employed techniques to mitigate dimensionality and address class imbalance. The synergistic application of these methodologies is intended to boost the predictive performance of the model. Analyzing the gene expression data is still challenging due to its many characteristics such as high dimensionality, complexity, and heterogeneity. DL algorithms have been applied and proven to be an effective technique for handling gene expression data, leading to significant improvement in the predictions and diagnosis of various types of cancer (Kosolwattana et al., 2023; Ravindran & Gunavathi, 2023; Mondol et al., 2023). The features are extracted from the original information provided by DL techniques. Various DL techniques are employed to examine the transcription sets. Convolution Neural Networks (CNN) is specifically used to train classifying parameters with several layers of kernel filters (Ravindran & Gunavathi, 2023; Mondol et al., 2023). CNN models have demonstrated outstanding classification performance in gene expression analysis because of their ability to capture local spatial relations from the input data. Therefore, CNN models have consistently classified among the top-performing DL models when applying gene expression data. A hybrid strategy based on Recurrent Neural Networks (RNNs) and CNNs is proposed (Thakur et al., 2023) for predicting various kinds of cancer using gene expression data. Measuring several metrics for

performance like accuracy, precision, recall, Mean Square Error (MSE), and F1 score, the suggested model outperforms all other current approaches, including VGG16, VGG19, ResNet50, Inception V3, and the mobile net classifier. For dataset 1, the RNN-CNN classifier yields an optimal precision of 0.978 compared to all the other techniques currently in use. At 80% of data used as training, it achieves the maximum reliability of 0.994 for Dataset 2. One can think of DL as a subset of ML. Due to its numerous benefits, including superior outcomes from DL techniques over ML models, readily accessible assets and information, and the constant development of new computational methods, DL is a rapidly expanding area in the modern period (Salas et al., 2019). The foundation of DL models is SGD. Despite the ML systems' excellent performance, they lag short in certain domains, such as objects and language recognition. This serves as one of the causes of the DL field (Goodfellow, 2016). DL concepts originated from ANNs that are impacted by the brain's neuronal activity. The input, hidden, and output layers—each with numerous nodes—are the three layers that make up an ANN. Regarding monitored tasks, generalization is absent in these simulations, which learn via a sequential learning procedure called backpropagation. A Deep Neural Network (DNN) is an ANN with additional hidden layers that offer superior generalization. Additionally, it facilitates enhanced extraction of characteristics and extensive parameter learning. Caffe, PyTorch, Theano, Tensor Flow, and additional systems are available for DL model implementation (Zhu et al., 2021). According to (Ang et al., 2015), unsupervised feature selection techniques are unreliable due to the potential for choosing insignificant characteristics or eliminating useful ones. According to (Hwang et al., 2018), a variety of single-cell RNA sequence technology allows for concurrent analysis of several hundreds of cellular transcriptomes in a particular specimen. This allows for the evaluation of variability in population, to uncover cell differentiation trajectories, identify intricate structures, disclose unusual cell populations, detect genes that exhibit variation between different cell types or between samples, and so on. An ANN model was presented (Sarra et al., 2023) to build a DL diagnostics tool for the prediction of cardiac disease. The accuracy of the created ANN prediction model was 93.44%, 7.5% superior to that of a conventional Support Vector Machine (SVM) model for ML. Furthermore, the learning and categorization times were lowered to just over thirty seconds by utilizing a more basic neural network.

3 Proposed Methodology

This section provides a thorough analysis of the suggested model. The three stages of the suggested approach are as follows: the first involves preprocessing using Standard Scaler; the second involves investigating the prospect of Uniform Manifold Approximation and Projection (UMAP) for DR in complex datasets like the expression of gene dataset; and the third involves building a unique ANN model to create a DL screening tool for the LUAD estimation of lung cancer gene classification. Figure 1 shows an overall design of the suggested scheme.

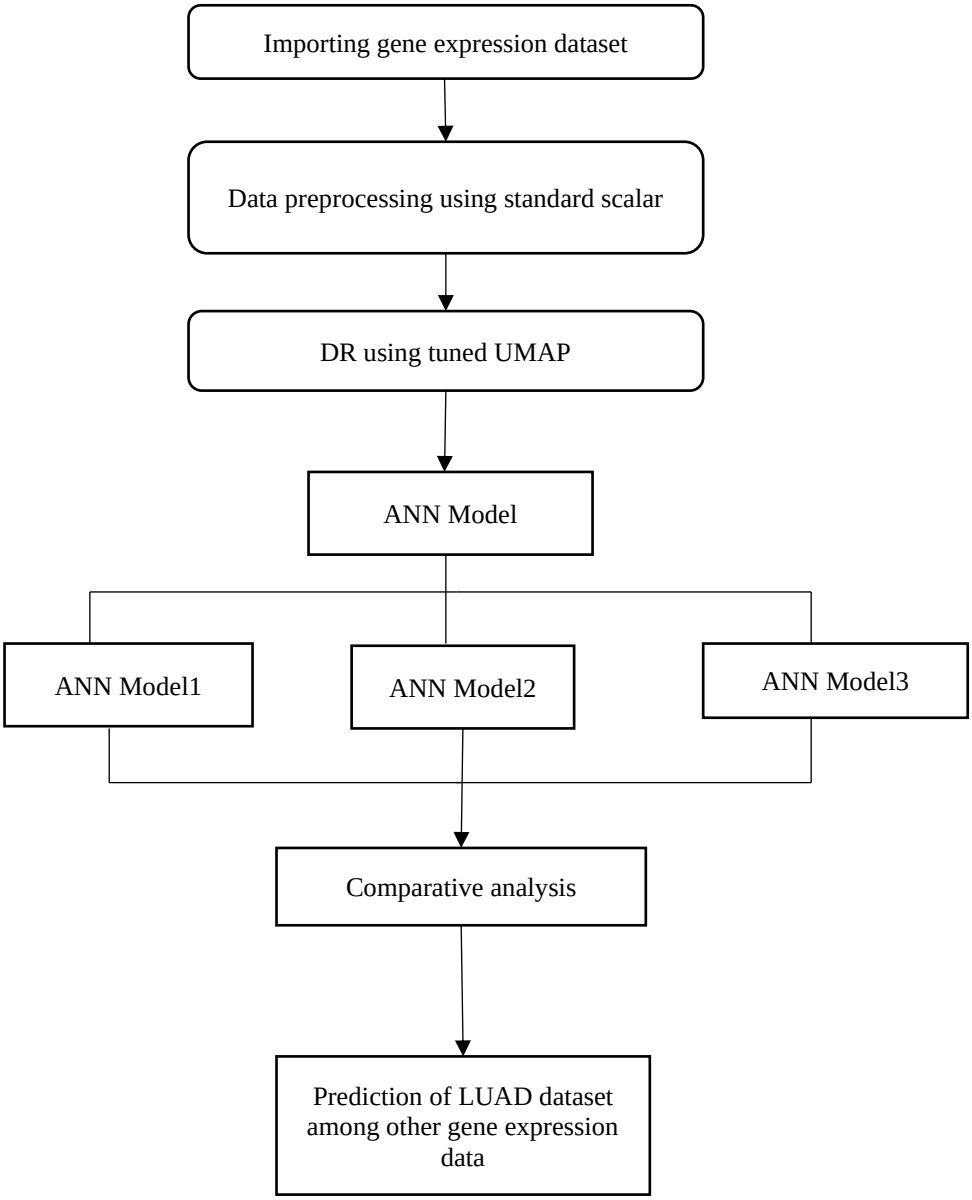

Figure 1: Overall Proposed Block Diagram

Dataset Description

The input dataset contains 20532 gene expressions and 5 samples are shown in Figure 2. Each sample contains expression values of more than 20K genes. Samples have one of the following kinds of tumour: PRAD, COAD, KIRC, LUAD and BRCA. It provides an erratic retrieval of gene expression from patients suffering from multiple tumours, such as Breast cancer or (BRCA), kidney renal cell carcinoma or (KIRC), colon adenocarcinoma or (COAD), LUAD and prostate adenocarcinoma (PRAD). From the various types of tumour, 802 gene expression is related to lung cancer (LUAD). A generic name in the form of "gene_XX" is assigned to each variable in the data file. The collection of data is accompanied with a CSV manifesto that provides record annotations and additional information for each file.

	Unnamed: 0	gene_0	gene_1	gene_2	gene_3	gene_4	gene_5	gene_6	gene_7	gene_8
0	sample_0	0.0	2.017209	3.265527	5.478487	10.431999	0.0	7.175175	0.591871	0.0
1	sample_1	0.0	0.592732	1.588421	7.586157	9.623011	0.0	6.816049	0.000000	0.0
2	sample_2	0.0	3.511759	4.327199	6.881787	9.870730	0.0	6.972130	0.452595	0.0
3	sample_3	0.0	3.663618	4.507649	6.659068	10.196184	0.0	7.843375	0.434882	0.0
4	sample_4	0.0	2.655741	2.821547	6.539454	9.738265	0.0	6.566967	0.360982	0.0

5 rows × 20532 columns

Figure 2: Sample Input Data Gene Expressions

Pre-processing Using Standard Scalar

The magnitude of the data must be normalized in order to avoid inaccurate results caused by input properties of gene data that may exist on various scales. Standard Scalar helps normalize the lung cancer database in this study. By deducting the mean value (referred as centering) and dividing by the standard deviation. To achieve a mean of zero and an acceptable variation of one, the system of standardization adjusts each of the input variables separately, modifying the way it is distributed. Sometimes the characteristics in an information set, vary greatly in their ranges or are measured in different units of measurement. The data are scaled to a variance of one after the mean is reduced to zero using the Standard scaler. However, the presence of data anomalies significantly affects the corresponding empirical mean and standard deviation calculations, narrowing the range of distinctive variables. The Standard Scaler function used by sklearn rests on the idea that variables in the set of data with values, fall into various ranges. They do not contribute equally to the fitted values and learning functionality of the model. This could potentially cause discrimination in the outcomes generated by the model.

Normalize attributes by eliminating mean values and scale them to unit variance. To find the average ratings of the sample (x), use the formula Z = (x-u)/s.

Here, s is the test samples standard deviation and u is the mean of zero of the reference sample, if with_mean = false, and one if with_std = false. Due to Standard scaler's sensitivity to anomalies, variables might measure differentially from one another when outliers are present.

DR by UMAP

In order to enable the 2- or 3-dimensional visualization of datasets with tens to hundreds of dimensions, DR techniques capture variability in a small number of random variables. The widely used PCA method, which creates orthogonal axes with fewer variables by combining linear combinations of variables, is recognizable as using this strategy to capture variance in data. In this research work proposed method utilizes tuned UMAP for preserving both local and global structures in high-dimensional data, and its integration in this study indicates a sophisticated approach to capturing intricate patterns within the gene expression data associated with lung cancer. Specifically, UMAP has proved helpful in accurately defining cell types in different populations using information from single-cell gene expression. It executes admirably on additional gold-standard datasets as well. Since data structure components from high-dimensional space may be preserved more effectively with UMAP.

The way of UMAP works is that it first constructs the weighted k-neighbourhood graph and then calculates its low-dimensional structure.

Let us consider the input dataset.

$D=\{d_1,d_2,\ldots,d_N\}$ with difference metric $g:D\times D\rightarrow r\geq 0$ for each d_i and an input hyperparameter k, we will compute ε_i and δ_i as follows in equation (1):

$$\delta_i=\min\{g(d_i,d_{ij}) \mid 1 \leq i \leq j,\ g(d_i,d_{ij}) > 0\} \quad (1)$$

ε_i represents set to satisfy the following in equation (2):

$$\sum_{i=1}^{j} \exp\left(\frac{-\max(0,d(d_i,d_{ij})-\delta_i)}{\varepsilon_{i_i}}\right)= \log_2(j) \quad (2)$$

Where:

$g(d_i,d_{ij})$ is k - nearest neighbor for respectively point d_i.

Then computing weighted directed graph $\bar{D}=(W,F,\omega)$, where the vertices W of $\bar{D}$ are the dataset G. The collection of directed edges $F=\{(d_i, d_{ij}) \mid 1 \leq i \leq j, 1 \leq J \leq M\}$. Equation (3) provides the computation of the weight function ω.

$$\omega((d_i, d_{ij}))= \exp\left(\frac{-\max(0,g(d_i,d_{ij})-\delta_i)}{\varepsilon_i}\right) \quad (3)$$

Equation (4) can be used to calculate the adjacency matrix of the undirected weighted graph A.

$$A=B+B^T-B\circ B^T \quad (4)$$

Where:

B is the weighted adjacency matrix of D–, and "∘" represents the Hadamard product.

In practice, UMAP utilizes a directed graph layout technique based on low-dimensional force.

This method exerts a gravitational force at the edges and a repulsive force at the vertex. Equation (5) can be used to calculate the gravitational force that exists between two vertices, i and j, with indices y_i and y_j.

$$\frac{-2ab\|y_i-y_j\|_2^{2(b-1)}}{1+\|y_i-y_j\|_2^2}\omega((d_i, d_{ij}))(y_i - y_j) \quad (5)$$

Where a and b represent the hyper-parameters.

The following Equation (6) is used to calculate the repulsive force.

$$\frac{2b}{\left(\epsilon+\|y_i-y_j\|_2^2\right)\left(1+a\|y_i-y_j\|_2^{2b}\right)}(1+\omega((d_i,d_{ij})))(y_i-y_j) \quad (6)$$

ϵ is a constant value that prevents division by zero.

After that class imbalance is a common issue where the distribution of examples across different classes is not equal. Therefore, certain classes have a lot fewer samples or instances than others. The vast majority is made up of all occurrences, whereas the minority classes are made up of every event less frequently. The Synthetic Minority Over-sampling methodology (SMOTE), which involves oversampling samples from the minority class to rebalance the gene expression data, is an efficient resampling methodology for unbalanced data classification. To resample the features, we define the target values as '1' and '0' where 1 specifies the identification of lung cancer-related genes as LUAD, 0 specifies other gene expressions such as BRCA, KIRC, COAD, and PRAD. Figure 3 illustrates the UMAP object for component values.

	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Comp8	Comp9
529	4.755271	2.359067	3.786187	4.404400	2.274770	7.380532	6.114050	4.734012	3.334993
310	7.500558	2.710219	5.122167	4.807616	3.037190	8.537114	4.122495	4.690825	4.273525
636	5.717943	2.263073	2.851655	4.460745	2.401554	7.926072	6.075662	4.379128	3.758663
620	5.434892	2.021389	3.565714	4.505230	2.799695	7.713995	6.164053	6.008351	3.446839
550	5.123683	2.789951	3.410607	4.023563	2.591151	7.847415	5.680648	4.711376	4.077606
...	...	...	...	...	...	...	...	...	...
645	6.116234	2.346406	4.103779	4.655691	2.536431	7.821635	5.608080	4.993882	3.828802
715	5.549674	2.302302	2.589896	4.256294	2.536346	7.871451	6.120337	4.699615	3.927005
72	7.245869	2.829581	4.649392	4.848642	2.803042	8.343149	4.591508	4.533472	4.121842
235	9.886667	5.231169	3.471610	5.996827	1.974048	7.551132	6.529187	2.796507	2.561496
37	0.952142	5.303739	2.332225	3.746465	1.848266	5.884255	6.901761	4.993566	3.790170

504 rows × 40 columns

Figure 3: UMAP Object for Trustworthiness Component Values

ANN Architecture

In order to increase diagnostic precision and forecast, along with gene expression data, LUAD kinds of cancer, ANN model is created and further optimized by modifying its hyperparameters. The numerical representation of biological neural pathways is provided by ANNs, which are controlled learning methods. The composition and operation of an ANN are modeled after those of the natural brain. A DL technique is comprised of a neural network that is artificial. A synthetic neural network served as the basis for DNN development. The total number of hidden layers is the only distinction between DNN and ANN; DNN has several hidden layers, while ANN only has one. The input layer, hidden layer, and output layer are the three separate layers that make up an ANN. The first layer receives inputs, and the final layer produces outputs, as seen in Figure 4.

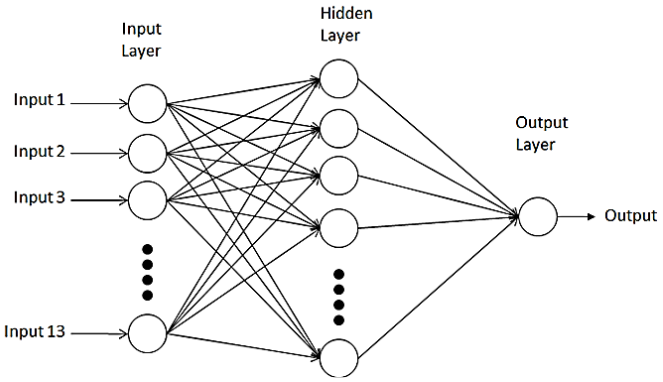

Figure 4: A Typical ANN Structure

The nodes that mimic neurons in the body are called artificial neurons, and they make up every level. All neuronal connections are given values. A back-propagation system that adjusts its weights is used for training artificial neural networks. The disparity in the expected and real results is taken into account when modifying the weights. Lastly, weight modifications are transmitted back to the source from the sink so that the input of the forward network may verify them. To provide outputs that are as similar to the target as possible, the process' ideal objective is to limit errors. The artificial neuron is the fundamental building block of an ANN (Flores-Fernandez et al., 2024).

The data points of every single neuron in each layer above, to which it is connected are added to determine the output of each neuron. By using an activation function, which is pre-defined and determined by the type of function, it may process the weighted inputs and output a number (often between 1 and -1) based on a threshold. Figure 5 illustrates the role of each neuron. To find out if an individual suffers from heart disorder or not, we created an ANN in this research. A pair of layers made up the model. The first input layer, which was the initial layer, included thirty units. The system received the heart-related characteristics of each patient through this input layer, where they were scaled by their corresponding weights. Then, in order to analyse the information that comes in x_i, nodes in the hidden layer compute the weighted total and add a bias b_i as mentioned in Equation (7). A weighted connection between nodes is represented by the symbol wij.

$$N_j = b_i + \sum_i^m x_i \, w_{ij} \tag{7}$$

Following that, N_j was transferred using the ELU-activated functioning, as shown in Equation (8):

$$y = \begin{cases} x \; when \; x \geq 0 \\ \alpha(e^x - 1) \; when \; x < 0 \end{cases} \tag{8}$$

Wherein α is a controllable variable that regulates the negative part of the ELU's saturation point. The final output was generated by a single node in the output layer using a sigmoid activation.

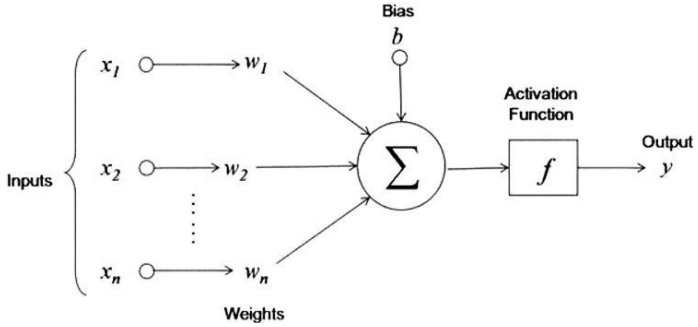

Figure 5: Neuron Components

Our network's connection weights were initialized at random to start the network process of training. The information that was provided was then analyzed by the neural network, which produced its result. Next, as shown in Equation (9), the system's outcome was contrasted with the intended output, and the resultant variance was computed using the method of binary cross-entropy loss.

$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}(y_i \times log\,(\hat{y}_i)) + (1 - y_i) \times log(1 - \hat{y}_i) \tag{9}$$

Where n is the total number of samples, ŷ and y stand for the expected and real outcomes, and L_{BCE} is the binary cross-entropy. The computed inaccuracy was transmitted through the computing system during learning, and parameters were changed as necessary, as seen in Equation (10).

$$\Delta w_{ij} = -\eta \frac{\partial Error}{\partial w_{ij}} \tag{10}$$

Where η, a constant that represents the relative weight change, and Δw_{ij} are the weight changes. The output value was recalculated following the updates to all of the weights based on the training faults. Until the network had converged with the fewest errors feasible, this procedure was continued.

To improve a neural network model's performance, its hyperparameters must be properly tuned. Hyperparameters are rigid, manually specified variables that are assigned during the modelling phase. Therefore, we used an optimization technique to optimize the hyperparameter combination before building our model. Using random and grid search approaches, the number of synapses in the layer that

is concealed, training rate, and activated function were all improved. Table 1 displays the ideal configuration of hyperparameters that was obtained through the use of optimization approach. To come up with different combinations of hyperparameter values, we first performed a hyperparameter search. Then we developed a model, one combination at a time. Following training and assessment, the average accuracy of the model was recorded. Subsequently, we duplicated the process with an alternative set of hyperparameter values. The perfect blend of hyperparameters for constructing our simulation was determined by taking into account every feasible choice and selecting the one that produced the greatest efficiency. The following manual modifications have been made to the other hyperparameters of the ANN model: 30 batches, which expedites the training procedure of 250 epochs for the model.

Table 1: Optimal Hyperparameter Tuning

ANN Model	Hyperparameter	Value
ANN Model 1	input layers	10 Neurons
	hidden layers	8 Neurons
	activated function	Relu
	rate of learning	0.001
	Output layer	SoftMax
ANN Model 2	input layers	20 Neurons
	hidden layers	10 Neurons
	activated function	Relu
	rate of learning	0.001
	Output layer	SoftMax
ANN Model 3	input layers	40 Neurons
	hidden layers	15 Neurons
	activated function	Relu
	rate of learning	0.001
	Output layer	SoftMax

4 Results and Discussion

The parametric form of UMAP yields embeddings that are comparable to those of the nonparametric form in this experimental study, including the advantage of a learned mapping between data space and embedding space. We used a few downstream activities to show how useful this learned mapping works. We demonstrated that, while retaining similar embedding quality to nonparametric UMAP, parametric relationships may be employed to increase inference speeds for embeddings and reconstructive by orders of magnitude. Parametric UMAP outperforms where the global structure is imposed only upon initialization, capturing additional global relationships in data when combined with a global structure preservation loss. The component value 8 attains the highest trustworthiness score of 0.9546 with a minimum distance of 0.9 is displayed in Figure 6.

```
UMAP_Object_8=umap.UMAP(n_components=40, n_neighbors=10, min_dist=0.9)
ComponentValues_8=UMAP_Object_8.fit_transform(X)
trustworthiness_score_8 = trustworthiness(X, ComponentValues_8)
print("Trustworthiness score of ComponentValues_8:", trustworthiness_score_8)
```

```
Trustworthiness score of ComponentValues_8: 0.9546845457158705
```

Figure 6: UMAP Component Value

Table 2 illustrates the comparison of a second nonparametric UMAP implementation that uses the same basic code as the parametric UMAP implementation, but instead of optimizing neural network weights, it optimizes over embeddings directly. In order to account for any potential implementation variations, this comparison serves as a bridge between the UMAP-learn implementation and Parametric UMAP. This compound sample is generated based on the feature extraction of UMAP algorithm and extracted features sample is implemented through proposed ANN model.

Table 2: Reduced Component Sample from Three Compounds for UMAP Analysis

	Comp1	Comp2	Comp3
0	5.150108	2.801036	3.265006
1	10.864498	5.456260	2.561375
2	7.664686	4.098872	2.762274
3	10.223743	5.426467	3.594932
4	-0.529161	5.713671	2.494600
5	11.083560	5.563601	2.920310
6	8.511126	4.347535	3.195442
7	8.244712	4.288273	2.836934
8	7.373535	4.487804	3.345193
9	5.019047	2.565576	3.695056

In this work, Google Colab is used with Jupiter IDE which assists in sharing and creating documents that can be narrated with text, live code and visualizations. The gene expression dataset is collected and split it into 70% train dataset and 30% test dataset. Figure 7 illustrates how the suggested ANN model is assessed in more detail using the Confusion Matrix (CM) metric. Figure 7 displays the CM that the ANN models 1 and 2 produced. The figure shows that the proposed ANN model1 and model2 can correctly detect 116 LUAD datasets and classify 98 out of 100 lung cancer gene expression data. The resulting CM of the ANN model3 is shown in the Figure 8. It demonstrates that 116 LUAD datasets were accurately recognized, and 99 of 100 lung cancer gene expression data were accurately classified. The target values as '1' and '0' where 1 specifies the identification of lung cancer-related genes as LUAD, 0 specifies other gene expressions such as BRCA, KIRC, COAD and PRAD.

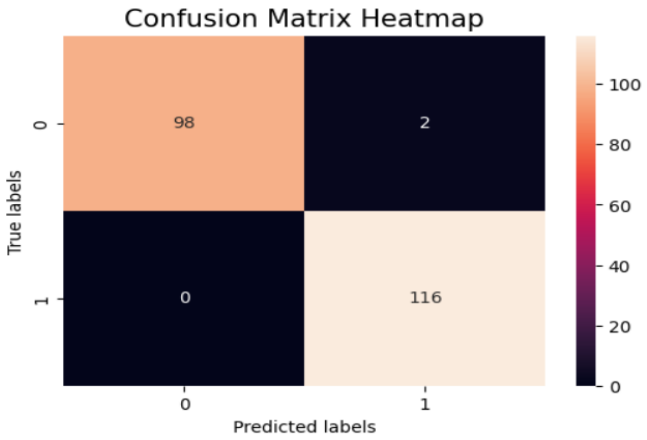

Figure 7: CM for ANN Model 1 and Model 2

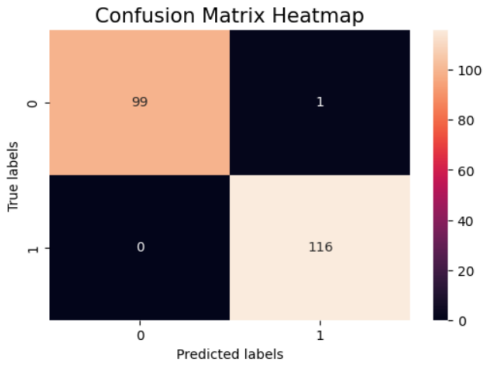

Figure 8: CM for ANN Model 3

Table 3 illustrates the performance measures using the CM which consists of actual and predicted classification information using the classification system. The research work used the python software to classify the dataset and its CM is also summarized. The CM of tunned ANN model 1 and model 2 in which True Negative (TN), False Positive (FP), False Negative (FN) and True Positive (TP) as a sequence value is 98, 2, 0 and 116 similarly ANN model3 the CM values of TN, TP, FN and TP as a sequence value is 99,1,0 and 116 correspondingly. Based on the values of interception among four different classes in CM, the efficiency of classification ANN model.

Table 3: CM Value for Comparison of Various ANN Models

S.No	Algorithm	CM Values			
		TP	TN	FP	FN
1	ANN model1	116	98	2	0
2	ANN model2	116	98	2	0
3	ANN model3	116	99	1	0

Table 4: A Comparative Analysis of ANN Model 1, 2 and 3 with their Performance Metrics

Performance Measures	ANN model1	ANN model2	ANN model3
Accuracy (%)	0.98	0.98	0.99
Precision	0.98	0.98	0.99
Recall	0.97	0.97	0.98
F1 score	0.98	0.98	0.99

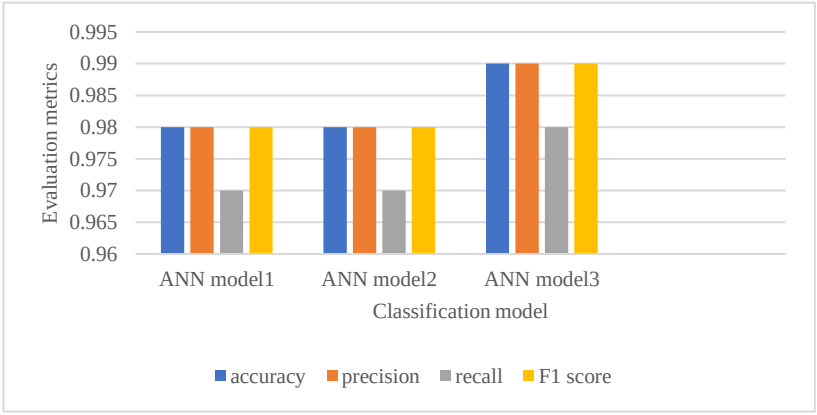

Figure 9: Performance Metrics for ANN Model1, ANN Model2 and ANN Model3

Table 4 presents the comparative results of ANN model1, 2 and 3. From the analysis, it is evident that the proposed ANN model3 has the highest accuracy of 0.99. From the analysis, Figure 9 has provide an evident that proposed ANN model 3 has better accuracy, precision and F1 score than the ANN model1 and ANN model 2.

5 Conclusion

Increasing diagnostic precision and identifying genes linked to lung cancer as LUAD are the two main objectives of this research. To achieve this purpose, an ANN model is built and made better by optimizing the parameters, which include the number of layers that are concealed, the count of axons in every layer, the pace at which it learns, and the function that activates them. Comparative analysis with recent competitive techniques demonstrated that our proposed method has an outstanding prediction performance, as indicated by various evaluation metrics. This indicates the potential of proposed ANN model to improve the diagnosis and understanding of LUAD, contributing significant insights to the domain of lung cancer-related research. It was clear that UMAP separated batching impacts, identified predetermined natural collections, and discovered clustered patterns related to biologic characteristics and clinically, by preserving the information of the sample in the community yet retaining precision. The highest F1 score and accuracy of proposed ANN model 3 attains 0.99 than ANN model 1 and ANN model 2. Future research in the domain of DL models for improved classification and prediction of lung cancer using 59890 multi-omics data could explore several promising directions to enhance the field.

References

[1] Alharbi, F., & Vakanski, A. (2023). Machine learning methods for cancer classification using gene expression data: A review. *Bioengineering*, *10*(2), 173. https://doi.org/10.3390/bioengineering10020173

[2] Alzaidi, E. R. (2024). Optimization of Deep Learning Models to Predict Lung Cancer Using Chest X-Ray Images. *International Academic Journal of Science and Engineering, 11*(1), 351–361. https://doi.org/10.9756/IAJSE/V11I1/IAJSE1140

[3] Ang, J. C., Mirzal, A., Haron, H., & Hamed, H. N. A. (2015). Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM transactions on computational biology and bioinformatics*, *13*(5), 971-989. https://doi.org/10.1109/TCBB.2015.2478454

[4] Archer, T. C., Fertig, E. J., Gosline, S. J., Hafner, M., Hughes, S. K., Joughin, B. A., ... & Shajahan-Haq, A. N. (2016). Systems approaches to cancer biology. *Cancer Research, 76*(23), 6774–6777. https://doi.org/10.1158/0008-5472.CAN-16-1580

[5] Bodor, J. N., Boumber, Y., & Borghaei, H. (2020). Biomarkers for immune checkpoint inhibition in non–small cell lung cancer (NSCLC). *Cancer, 126*(2), 260-270. https://doi.org/10.1002/cncr.32468

[6] Cai, L., Lin, S., Girard, L., Zhou, Y., Yang, L., Ci, B., ... & Xie, Y. (2019). LCE: an open web portal to explore gene expression and clinical associations in lung cancer. *Oncogene*, *38*(14), 2551-2564. https://doi.org/10.1038/s41388-018-0588-2

[7] Filipp, F. V. (2017). Precision medicine driven by cancer systems biology. *Cancer and Metastasis Reviews*, *36*, 91-108. https://doi.org/10.1007/s10555-017-9662-4

[8] Flores-Fernandez, G. A., Jimenez-Carrion, M., Gutierrez, F., & Sanchez-Ancajima, R. A. (2024). Genetic Algorithm and LSTM Artificial Neural Network for Investment Portfolio Optimization. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 15*(2), 27-46. https://doi.org/10.58346/JOWUA.2024.I2.003

[9] Ginn, L., Shi, L., La Montagna, M., & Garofalo, M. (2020). LncRNAs in non-small-cell lung cancer. *Non-coding RNA*, *6*(3), 25. https://doi.org/10.3390/ncrna6030025

[10] Goodfellow, I. (2016). Deep Learning. I. Goodfellow, Y. Bengio, A. Courville.

[11] Huang, K., Zhang, Y., Shi, X., Yin, Z., Zhao, W., Huang, L., ... & Zhou, X. (2023). Cell-type-specific alternative polyadenylation promotes oncogenic gene expression in non-small cell lung cancer progression. *Molecular Therapy-Nucleic Acids*, *33*, 816-831. https://doi.org/10.1016/j.omtn.2023.08.005

[12] Hwang, B., Lee, J. H., & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, *50*(8), 1-14. https://doi.org/10.1038/s12276-018-0071-8

[13] Khalsan, M., Machado, L. R., Al-Shamery, E. S., Ajit, S., Anthony, K., Mu, M., & Agyeman, M. O. (2022). A survey of machine learning approaches applied to gene expression analysis for cancer prediction. *IEEE Access*, *10*, 27522-27534. https://doi.org/10.1109/ACCESS.2022.3146312

[14] Kosolwattana, T., Liu, C., Hu, R., Han, S., Chen, H., & Lin, Y. (2023). A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare. *BioData Mining*, *16*(1), 15. https://doi.org/10.1186/s13040-023-00330-4

[15] Liu, S., Xu, C., Zhang, Y., Liu, J., Yu, B., Liu, X., & Dehmer, M. (2018). Feature selection of gene expression data for cancer classification using double RBF-kernels. *BMC bioinformatics*, *19*, 396. https://doi.org/10.1186/s12859-018-2400-2

[16] Mathew, C., & Asha, P. (2024). FedProx: FedSplit Algorithm based Federated Learning for Statistical and System Heterogeneity in Medical Data Communication. *Journal of Internet Services and Information Security, 14*(3), 353-370. https://doi.org/10.58346/JISIS.2024.I3.021

[17] Mohamed, T. I., & Ezugwu, A. E. (2024). Enhancing Lung Cancer Classification and Prediction with Deep Learning and Multi-Omics Data. *IEEE Access*, *12*, 59880 – 59892. https://doi.org/10.1109/ACCESS.2024.3394030

[18] Mondol, R. K., Millar, E. K., Graham, P. H., Browne, L., Sowmya, A., & Meijering, E. (2023). hist2rna: an efficient deep learning architecture to predict gene expression from breast cancer histopathology images. *Cancers*, *15*(9), 2569. https://doi.org/10.3390/cancers15092569

[19] Organization. (2023). W.H. Lung Cancer. https://www.who.int/news-room/fact-sheets/detail/lung-cancer

[20] Osama, S., Shaban, H., & Ali, A. A. (2023). Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review. *Expert Systems with Applications*, *213*, 118946. https://doi.org/10.1016/j.eswa.2022.118946

[21] Ramakrishnan, J., Ravi Sankar, G., & Thavamani, K. (2019). Publication Growth and Research in India on Lung Cancer Literature: A Bibliometric Study. *Indian Journal of Information Sources and Services, 9*(S1), 44–47. https://doi.org/10.51983/ijiss.2019.9.S1.566

[22] Ravindran, U., & Gunavathi, C. (2023). A survey on gene expression data analysis using deep learning methods for cancer diagnosis. *Progress in Biophysics and Molecular Biology*, *177*, 1-13. https://doi.org/10.1016/j.pbiomolbio.2022.08.004

[23] Salas, J., de Barros Vidal, F., & Martínez-Trinidad, F. (2019). Deep learning: current state. *IEEE Latin America Transactions*, *17*(12), 1925-1945. https://doi.org/10.1109/TLA.2019.9011537

[24] Saritha, R. R., & Gunasundari, R. (2024). Enhanced Transformer-based Deep Kernel Fused Self Attention Model for Lung Nodule Segmentation and Classification. *Archives for Technical Sciences, 2*(31), 175–191. https://doi.org/10.70102/afts.2024.1631.175

[25] Sarra, R. R., Dinar, A. M., & Mohammed, M. A. (2023). Enhanced accuracy for heart disease prediction using artificial neural network. *Indonesian Journal of Electrical Engineering and Computer Science, 29*(1), 375-383., https://doi.org/10.11591/ijeecs.v29.i1.pp375-383

[26] Thakur, T., Batra, I., Luthra, M., Vimal, S., Dhiman, G., Malik, A., & Shabaz, M. (2021). [Retracted] Gene Expression-Assisted Cancer Prediction Techniques. *Journal of Healthcare Engineering*, *2021*(1), 4242646. https://doi.org/10.1155/2021/4242646

[27] Thakur, T., Batra, I., Malik, A., Ghimire, D., Kim, S. H., & Hosen, A. S. (2023). RNN-CNN based cancer prediction model for gene expression. *IEEE Access*, *11*, 131024-131044. https://doi.org https://doi.org/10.1109/ACCESS.2023.3332479

[28] Vos, T., et al., (2016). Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: A systematic analysis for the global burden of disease study 2015. *Lancet*, *388*(10053), 1545–1602. https://doi.org/10.1016/S0140-6736(16)31678-6

[29] Xue, Q., Peng, W., Zhang, S., Wei, X., Ye, L., Wang, Z., ... & Zhou, Q. (2023). Promising immunotherapeutic targets in lung cancer based on single-cell RNA sequencing. *Frontiers in Immunology*, *14*, 1148061. https://doi.org/10.3389/fimmu.2023.1148061

[30] Yuan, F., Lu, L., & Zou, Q. (2020). Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, *1866*(8), 165822. https://doi.org/10.1016/j.bbadis.2020.165822

[31] Zhu, T., Li, K., Herrero, P., & Georgiou, P. (2021). DL for diabetes: A systematic review. IEEE Journal of Biomedical and Health Informatics, 25(7), 2744–2757.

Authors Biography

B. Jyothi, is a graduate with M.Tech in Computer Science Engineering from SRM Institute Of Science And Technology, Chennai India. She is currently pursuing her Ph.D. degree in Computer Science engineering at Sathyabama Institute of Science & Technology. She is currently working as an Assistant Professor at Computer Science Engineering and AIML department in Malla Reddy College of Engineering and Technology, Hyderabad, India. Her research interests are in fields of Computer Vision, Machine Learning.

Dr.L. Mary Gladence, is a Professor in the School of Computing, Sathyabama Institute of Science and Technology, Chennai, India. She is having 18 years of experience. Her research interests include deep learning, artificial intelligence, data mining, sequential pattern Mining, machine learning, bio computing, data analytics, with more than 70 publications in these areas. She has been a guest editor and reviewer in refereed international journals like, IEEE Access, Library Hi-Tech, Journal of Super Computing, Journal of Medical and Biological Engineering, Computer communications, etc. She has completed a project titled “Cattle form management using RFID tags”, funded by Unnat Bharat Abhiyan (UBA), MHRD, Govt of India and currently working on the project titled “Inventorization of Waste Management - The Global Scenario”, Technology Business Incubator, NSTEDB - DST, Govt of India.