Ensemble Deep Learning based Lung Cancer Classification Model using Gene Expression Data

V. Yuvaraj [1*], D. Maheswari [2]

Abstract

Background: Globally, lung cancer is the deadliest form of the disease. Genetic variability is one of the elements that influence an individual's vulnerability to lung cancer, according to epidemiological research. Asian women, smokers or not, have a higher risk of acquiring cancer because of genetic abnormalities, according to a recent investigation from the US National Cancer Institute that involved 14,000 Asian women. A superior approach for classifying lung cancer was presented in recent studies to address the aforementioned issue. In this study, the data scale is first normalized utilizing min max normalization, which is accomplished by data pre-processing. Methods: Gene selection is carried through employing Improved Whale Optimization Algorithm (IWOA). An Enhanced Convolutional Neural Network (ECNN) is employed for lung cancer categorization. However, lung cancer classification using single algorithm produces insufficient accuracy. This required the need for development of ensemble models. To evade this issue,input data scales are normalized based on Z score normalization model. Once the normalization is done, significant genes are selected from these normalized gene samples using Modified Chicken Swarm Optimization (MCSO). Results: Finally, ensemble of ECNN, VGG16 and ResNet50 models are employed for lung cancer classification. Ensemble learning is performed in this work using majority voting. Conclusion: The suggested approach outperforms various alternatives in the field of accuracy, according to the findings.

Keywords: Lung cancer, Microarray analysis, Gene selection, Ensemble learning, Deep learning

Significance | Precision medicine, microarray analysis, ensemble learning, and deep neural networks revolutionize lung cancer diagnosis, enhancing accuracy and treatment efficacy.

*Correspondence. V. Yuvaraj, Department of Computer Science, RVS College of Arts and Science, Coimbatore 641402, Tamil Nadu, India.
E-mail: yuvarajvelliangiri@gmail.com

Editor Surendar Aravindhan, And accepted by the Editorial Board Apr 04, 2024 (received for review Feb 28, 2024)

Introduction

Lung cancer (LC) death rates are rising in the modern era, based on data from several health organizations (Salem et al., 2017). There has been a substantial increase in the death ratio from cancer since the turn of the decade. Even with the widespread use of therapies such as radiation, chemotherapy, and surgery, more work needs to be done to achieve favorable outcomes. Precisely categorizing various forms of lung cancer is vital for optimizing treatment efficacy and minimizing harmful effects on individuals (Diaz et al., 2014).

Microarray analysis enables the study of millions of genes, providing essential data about cellular functions (Dass et al., 2014). This important data can significantly impact the prognosis and treatment of cancer. Given the features of gene expression data, it is crucial to develop an excellent strategy for identifying significant gene subsets that can be leveraged for more accurate cancer categorization. Utilizing this strategy, medical professionals can focus on specific genes and create less costly studies by classifying a

Author Affiliation.
[1] Department of Computer Science, RVS College of Arts and Science, Coimbatore 641402, Tamil Nadu, India
[2] Department of Computer Science, RVS College of Arts and Science, Coimbatore 641402, Tamil Nadu, India.

Please cite this article.
V. Yuvaraj, D. Maheswari. (2024). Ensemble Deep Learning based Lung Cancer Classification Model using Gene Expression Data, Journal of Angiotherapy, 8(4), 1-9, 9636

smaller selection of biologically related tumors characterized by their genes. This approach also helps reduce analytical expenses. Additionally, this highly accurate classification method aids in medication discovery and early identification of cancer patients (Almugren et al., 2019; Cahyaningrum and Astuti, 2020).

Furthermore, prior knowledge is applied to verify the reliability of existing knowledge and to validate experimental data, adding new information or closing any gaps. Different gene expression profiles have been precisely categorized across tumor subtypes recently (Yuan et al., 2020; Azzawi et al., 2017). Studies have demonstrated that judicious use of readily available biological data can prevent biased outcomes from individual experiments and successfully eliminate noise in gene chips (Haznedar et al., 2021; Venkatesan et al., 2022). However, the potential of historical data for classifying cancers has not yet been fully acknowledged.

A superior approach for classifying lung cancer was presented in current research to address the aforementioned issue. In this study, data scale normalization is first achieved using min-max normalization during data preprocessing. Gene selection is performed using the Improved Whale Optimization Algorithm (IWOA). The Enhanced Convolutional Neural Network (ECNN) is then employed for lung cancer categorization. However, lung cancer classification using a single algorithm often produces insufficient accuracy, necessitating the development of ensemble models.

The study provided a system for the categorization of lung cancer to address classification issues. In the preprocessing phase, input data scales are normalized using the Z-score normalization model. During the gene selection phase, significant genes are selected from these normalized gene samples using the Modified Cuckoo Search Optimization (MCSO), which avoids local optima through the use of a mutation operator. In the detection phase, a majority voting-based ensemble of Enhanced Convolutional Neural Network (ECNN), VGG16, and ResNet50 models is employed for lung cancer classification. This comprehensive approach improves the accuracy and reliability of lung cancer categorization.

Literature Review

Hu et al. (2019) suggested a system that classifies lung adenocarcinoma (LUAD) into four subtypes. Five genes might serve as LUAD markers, while 24 differentially expressed genes can be exploited as treatment targets. The subtypes function as prognostic subtypes according to a multivariate approach. Targetable indicators for the various subtypes were identified by analyzing relevant genes. The function and pathway enrichment analysis of these representative genes revealed that the four subtypes have distinct pathogenic pathways. Drug development might consider subtype-related mutations as possible indicators; subtypes 1 and 2 have TP53 mutations, while subtype 4 has EGFR mutations. These four subtypes serve as a basis for LUAD subtype-specific therapy.

Azzawi et al. (2019) proposed a method based on the MLP-IMPSO technology. To improve classification accuracy, the approach incorporates lung cancer categorization based on Gene Expression Data (GED). Utilizing actual microarray lung cancer datasets, extensive assessments and evaluations of prediction accuracy were conducted among the proposed approach and relevant machine learning techniques. The evaluation was deemed trustworthy due to cross-dataset validations. After previous information was incorporated, the proposed strategy performed superiorly. The findings demonstrated the efficacy of the proposed method for diagnosing lung cancer.

Arunkumar and Ramakrishnan (2018) suggested a fuzzy rough quick reduct approach that establishes a personalized resemblance metric for selecting the minimum number of useful genes. Leukemia, lung, and ovarian cancer (OC) gene expression (GE) datasets were utilized to assess the proposed approach using a Random Forest (RF) classifier. The lung, leukemia, and OC GE datasets achieved classifier accuracies of 99.45%, 97.22%, and 99.6%, respectively, with the proposed approach. Compared to current techniques, the proposed approach performs better in terms of f-measure, recall, accuracy, and precision in classification.

Azzawi et al. (2018) suggested a novel Sample-Based Clustering (SBC) method for exploiting microarray data to identify subgroups of lung cancer. Gene Expression Profiling (GEP) is the foundation of the strategy. Extensive classification efficacy assessments and analyses were carried out using real microarray lung cancer datasets. These assessments compared the GEP system with popular binary decomposition methodologies and three techniques: Support Vector Machine (SVM), neural network, and C4.5. The reliability of cross-dataset validation was determined. According to the findings, the proposed approach outperformed other methods in terms of accuracy, standard deviation, and Area Under the Curve (AUC).

Jinthanasatian et al. (2017) introduced a neuro-fuzzy firefly system applied to microarray categorization. This system uses a neuro-fuzzy classifier to produce rule sets and select appropriate feature sets. The outcomes from seven public datasets, including lung cancer (LC), ovarian cancer (OC), acute lymphoblastic leukemia (ALL), colon cancer, and diffuse large B-cell lymphoma (DLBCL), were evaluated against current methods. It was discovered that the neuro-fuzzy firefly system could achieve comparable results with fewer selected features.

Wang et al. (2018) suggested a Weighted Group Graphical Lasso (WGGL) framework for grouping cancer genes. The framework is based on weighted gene co-expression network evaluation and employs a heuristic approach for gene grouping. It includes a technique for estimating gene and group weights based on joint

shared data to assess the relative relevance of genes and groups. A gene selection method was developed to manage the complex computation process of WGGL. According to the findings on three cancer gene expression datasets and a random dataset, the proposed approach outperforms two contemporary gene selection techniques in terms of classification accuracy.

Chaudhari and Agarwal (2018) introduced a feature set selection method for cancer categorization using microarray gene expression data. Due to the imbalance in the number of samples and genes, studying feature selection methods from complex gene expression data is crucial. They conducted a study on gene datasets using Elitist Binary Quantum Particle Swarm Optimization (EBQPSO). The findings demonstrate that the EBQPSO method, which integrates Particle Swarm Optimization (PSO) and Quantum PSO (QPSO), improves accuracy and recall value in deep searching and categorization of genetic datasets.

Methodology

The recommended framework is enclosed extensively in this section. There are three stages to the suggested approach. First one is Z score normalization based data normalization, second one is gene selection using modified chicken swarm optimization and the third one is classification using ensemble of ECNN, VGG16 and ResNet50 models. Figure 1 illustrates the overall structure of the recommended paradigm.

Data normalization (DN) with Z score normalization (ZSN)

Normalization is required to regulate the input scale after data balance. The common DM (Data Mining) steps to increase the precision of machine learning (ML) method is data pre-processing. Normalization is applied to all the data before to training and testing. This is employed to guarantee that data is not overloaded with one another. Data from several scales are converted to the same scale utilizing the normalizing process. This Z-score normalization process employs feature A's mean and standard deviation to normalize values. The following formula is applied:

$$v' = \frac{v - \bar{A}}{\sigma_A} \quad (1)$$

Where,

v', v - each data entry's new and old, accordingly

$\bar{A}$, σ_A - the mean of A and its standard deviation, accordingly

Gene selection using modified chicken swarm optimization

After normalize the data, it required to select important features from the database, for which here used modifiedCSO.

Chicken Swarm Optimization (CSO)

The CSO method emulates the individual chickens' behaviors along with the hierarchical structure of of chickens (Hafez et al., 2015); (Tripathi et al., 2020). A chicken swarm's structure is separated into multiple categories, each of which has a rooster and numerous hens and chicks. The laws of motion that apply to various kinds of chickens vary. Chickens' social lives are significantly influenced by a system of hierarchy. A flock of hens will be dominated by its stronger members (Zarlis et al., 2016). Both the subservient hens and roosters who gather at the group's boundaries and the more powerful hens who stay close to the head roosters are present(Moldovan 2020).

Traditional CSO will easily falls into the trap of local optimal features.

To avoid this problem this work used mutation operator in CSO. This work used flip bit mutation. This operator for mutation accepts the selected genome and flips its bits. (The genomic bit switches from 1 to 0 and vice versa if it is a 1).

Modified Chicken Swarm Optimization (MCSO)

The MCSO structure was recommended in with assistance of the subsequent rules, that perfectly capture the actions of the hens.

1) There are numerous groupings within the swarm of chickens. Usually is a leading rooster in each group, followed by a few hens and chicks.

2) The roosters, who have the best fitness values, leading the flock, whereas the distinct birds are the chicks, which have the least fitness values. The flock's ordering depends on the fitness scores of the chickens.

3) In a team, the mother-child bond, leadership dynamic, and swarm hierarchy won't alter. Only some (G) time steps pass amongst updates to these statuses.

4) The N virtual chickens that constitute up the swarm are separated into the sets: RN, CN, HN, and MN, which stand for the quantity of roosters, chicks, hens, and mother hens, accordingly. Positions of each person in a D-dimensional space are portrayed as

$$x_{i,j(i\in[i,\ldots\ldots N],j\in[1,\ldots\ldots,D])}, \quad (2)$$

Rooster Movement: Equations (3) and (4) illustrate why roosters with higher fitness values may look for food in a greater variety of locations than those with fewer fitness values.

$$x_{i,j}^{t+1} = x_{i,j}^{t} * \left(1 + Randn(0,\sigma^2)\right) \quad (3)$$

$$\sigma^2 = \begin{cases} 1, & if f_i \le f_k, \\ exp\left(\frac{f_k - f_i}{|f_i| + \epsilon}\right) & otherwise k \in [1,N], k \ne i, \end{cases} \quad (4)$$

Where as $x_{i,j}$ is the particular rooster using index i, standard deviation σ^2 , Rand $n(0,\sigma^2)$is a Gaussian distribution through mean 0.$\in$ $s$the smallest computer constant that prevents zero-division error, k is a arbitrarily generated rooster index taken from the roosters company, and f_i is the associated rooster x_i's fitness value.

Hen movement: A bunch of hens look for food by trailing behind roosters. In addition, while being suppressed by the remaining chickens, they would haphazardly steal the tasty food that they discovered. In a competition for food, the more assertive chickens

would have a benefit over the more timid ones. Equations (5) and (6) provide a formulation for these instances.

$$x_{i,j}^{t+1} = x_{i,j}^{t} + S1 * rRand * \left(x_{r_{1,j}}^{t} - x_{i,j}^{t}\right) + S2 * Rand * \left(x_{r_{2,j}}^{t} - x_{i,j_{\square}}^{t}\right) \quad (5)$$

$$S1 = exp\big((f_i - f_{r1})/ abs(f_i)+\in\big)) \quad (6)$$

$$S1 = exp\left(\left(f_{r_{2-\square}} f_i\right)\right) \quad (7)$$

Whereas, A uniform random number across [0, 1] is called a Rand. $r_1 \in$ [1,….., N] is rooster's index, that is i^{th} hen's group-mate, whereas $r_2 \in$ [1,….., N, is arbitrarilyselectedchicken index from the swarm.

Chick movement: The chicks follow their mother throughout in an attempt to get nourishment. It is expressed in equation (8).

$$x_{i,j}^{t+1} = x_{i,j}^{t} + FL * \left(x_{m,j}^{t} - x_{i,j}^{t}\right) \quad (8)$$

Where $x_{m,j}^{t}$ is the spot of the mother of the ith chick which m ∈ [1;N], The FL variable specifies the variations amongst each chick and indicates the pace at which a chick is following its mother. FL is selected at random from the interval [0, 2].

To identify the optimum location in the search space that optimizes the specified FF, one must employ a smart finding methodsince the feature space is so huge, where each feature is depicted by a separate dimension with a span of 0 to 1. Equation (9) illustrates the FF, which is to optimize the classification accuracy across the validation set provided the training data while maintaining the smallest amount of selected features.

$$f_\theta = \omega * E + (1 - \omega)\frac{\sum_i \theta_i}{N} \quad (9)$$

Where as N is the total quantity of features in the dataset, E is the classifier error rate, and ω is a constant controlling the significance of the classifier's accuracy to the quantity of features chosen. Given a vector θ with 0/1 elements indicating unselected / selected attributes, f_θ is the (FF) fitness function.

The quantity of features in the provided dataset matches the quantity of parameters utilized. Each parameter is constrained to the interval [0, 1], whereby its value reaches 1 and the associated feature is a potential candidate for classification selection. The factor in the individual fitness analysis is the threshold, which determines which qualities in particular need to be assessed according to equation (10).

$$f_{i,j} = \begin{cases} 1 \; if X_{i,j} > 0.5 \\ 0 \; otherwise, \end{cases} \quad (10)$$

Where in search agent i's dimension value at dimension j is represented by X_{ij}. To safeguard variable limitations when updating the firefly position solution, a simple truncation rule was employed because the updated value may breach the maximum constraints at certain dimensions [0, 1].

1. Set RN, HN, CN, MN, G;
2. Start every chicken in the swarm at arbitrarily.
3. X_i (i = 1,2,……;N).;
4. Set the max numbers of iteration T_{max};
5. while T <T_{max} do for every iteration
6. if T % G equals 0 then
7. Sort the hens according to their fitness levels and create a hierarchy inside the swarm;
8. Split the swarm into several sets and ascertain how each group's chicks and mother hens interact.
9. end
10. for every chicken Xi in the swarm do
11. if Xi is a roster then
12. Update Xi's location utilizing equation 8;
13. end
14. if Xi is a hen then
15. Update Xi's location utilizing equation 5;
16. end
17. if Xi is a chick then
18. Update Xi's location utilizing equation 8;
19. end
20. Assess the novel solution utilizing equation 10;
21. If the new solution is better than its previous one, update it;
22. end
23. end
24. Apply flip bit mutation to the updated solution
25. Estimate the novel solution utilizing equation 10
26. end

Classification using ensemble of ECNN, VGG16 and ResNet50 models

Following feature selection, an ensemble of ECNN, VGG16, and ResNet50 models are employed to classify lung cancer.

Enhanced CNN

Three different types of layers comprise up the CNN: fully linked, subsampling, and convolution layers. Figure 2 depicts an ordinary CNN design.

Convolution layer

A kernel (filter) is employed in this convolution layer to gather an input feature (Albahar 2019). *n* output feature maps are produced utilizing the input feature and kernel convolution output. The output features produced by assembling the kernel and the input are termed as FM (Feature Maps) of size i*i, but the kernel of the convolution matrix is usually termed as a filter.

Fuzzy Membership Function (FMF), that is utilized for FM, is described as $(w_1 = 0.3, w_2 = 0.4, w_3 = 0.5, w_4 = 0.7)$ and considered as

$$\boldsymbol{o^2} = \boldsymbol{u_i^{(j)}}\left(\boldsymbol{a_i^{(2)}}\right) \quad (11)$$

Where $\boldsymbol{u_i^{(j)}}(.)$ is a membership function$\boldsymbol{u_i^{(j)}}(.): \boldsymbol{R} \rightarrow [\boldsymbol{0}, \boldsymbol{1}]$, i=1,2,…,M, j =1,2,….,N. By utilizing the (GMF) Gaussian membership function.

The output of the *l*-th convolution layer, denoted as $C_i^{(l)}$, consists of FM calculated as

$$C_i^{(l)} = B_i^{(l)} + \sum_{j=1}^{a_i^{(l-1)}} K_{i,j}^{(l-1)} * C_j^{(l)} \quad (12)$$

Whereas, $B_i^{(l)}$is the bias matrix and $K_{i,j}^{(l-1)}$is convolution filter or kernel of size a^*a that connectsthe i-th FM in a similar layer as the j-th FM in layer $(l-1)$.

The output $C_i^{(l)}$ layer contains FM. In (12), the first convolutional layer $C_i^{(l-1)}$is input space, that is, $C_i^{(0)} = X_i$. A FM is created by the kernel. The activation function utilized to perform a nonlinear modification of the convolutional layer's outputs afterward the convolution layer.

$$Y_i^{(l)} = Y(C_i^{(l)}) \quad (13)$$

Where, $Y_i^{(l)}$ is output of the AF (Activation Function) and $C_i^{(l)}$is the data that it gets.

Sub sampling or pooling Layer

Decreasing the size of the FM that was adopted by the prior convolution layer geographically is the primary goal of this layer. Among the mask and the FM, a subsampling procedure is carried out. Numerous subsampling techniques were put forth, including maximum, sum, and average pooling. The max pooling, in each block's maximum value links to an output feature. Recall that the convolution layer can withstand rotation and translation amongst the input data with the assistance of a subsampling layer.

Fully Connected layer (FC) layer

A conventional feed forward network containing one or more hidden layers assists the last layer of a CNN. The AF Softmax is utilized in the output layer:

$$Y_i^{(l)} = f(z_i^{(l)}), \quad (14)$$

$$\text{Where } z_i^{(l)} = \sum_{i=1}^{m_i^{(l-1)}} w_{i,j}^{(l)} y_i^{(l-1)} \quad (15)$$

Where, $w_{i,j}^{(l)}$ are the weights for every class's image must have formed by fine-tuning the entire fully linked layer, and f is the transfer function specifies nonlinearity. Observe that unlike convolutions and pooling layers, which have nonlinearity constructed in separate layers, the FC layer has nonlinearity incorporated within the neurons.

VGG-16 model

Figure 3 illustrates the VGG16 framework, which is an CNN framework with 13 convolutional layers(CL) (C1, C2, C3 to C13) and 3 FC layers (FC-6, FC-7, and FC-8). Only three x three kernel sizes are employed in the VGG16 network's architecture, with each CL floating over top of the others to enhance depth(Kaur and Gandhi 2019). A stack of CL with a 3 x 3 kernel size was sent through the initial CL feed input in the pre-trained framework. Following the max-pooling layer are some convolutional layers with a 2 x 2 filter size. Three FC layers with varying depths and formats follow the set up of all CL. The first two entirely linked layers have 4096 channels, and the third FC layer has 1000 channels since it uses 1000-way ILSVRC categorization.

ResNet50

As seen in picture 4, ResNet-50 is a ResNet variation with 50 layers. Three types of layers were processed by ResNet-50: 48 CL, 1 MaxPool layer, and 1 average pool layer(Al-Haija and Adebanjo 2020). The fundamental idea behind ResNets is to employ shortcuts to get around CL bottlenecks. The fundamental Block, known as the "bottleneck," adheres to two important design principles: For a given output FM size, layers have an equal number of filters; if the FM size is half, the quantity of filters is doubled (Tian and Chen 2019); (Metwalli et al., 2020). Figure 3 illustrate the ResNet-50 structure.

Ensemble learning

Majority Voting (MV) is a technique for making decisions that is derived from classifiers which are run n times, independently, and separately, each time providing additional capabilities. Let C be an array of Q classes and χ be a collection of N examples. Defining a method set $S = \{A_1, A_2, A_M\}$, comprising the M classifiers utilized in the voting process, is necessary. The Q classes is allocated to each case $x \in \chi$. Every instance will have a forecast for every time classifier. Each sample's final class is the one that the vast majority of classifiers projected for this particular case.Every vote in MV is weighted according to the classifier's accurate prediction value, represented by the letter Acc. Then, the total number of votes for a class c_k expressed as follows:

$$T_k = \sum_{l=1}^{M} Acc\ (A_l) \times F_k(c_l) \quad (16)$$

$$F_k(c_l) = \begin{cases} 1 & c_l = c_k \\ 0 & c_l \neq c_k \end{cases} \quad (17)$$

Where c_l and c_k arethe classes of C. The class with the highest cumulative weight is selected. Essentially, weights are allocated to each classifier after it has been trained on various independent training sets, resulting in the maximum classification rate possible for classifying the data as positive or negative.

Performance Metrics

The ratio of accurately discovered positive outcomes to all anticipated positive data is recognized as precision.

$$\text{Precision} = TP/TP+FP \quad (18)$$

The F1 score is defined as the weighted average of the Precision and Recall. It involves false positives and false negatives as an outcome.

$$\text{F1 Score} = 2^*(\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}) \quad (19)$$

The following is the calculation of accuracy in positives and negatives:

$$\text{Accuracy} = (TP+FP)/(TP+TN+FP+FN) \quad (20)$$

Specificity quantifies the percentage of actual negatives that the framework accurately detects in the manner described below.

$$\text{Specificity} = TN / TN + FP \quad (21)$$

While TN- true negative, TP –true positive, FN –false negative and FP –false positive.

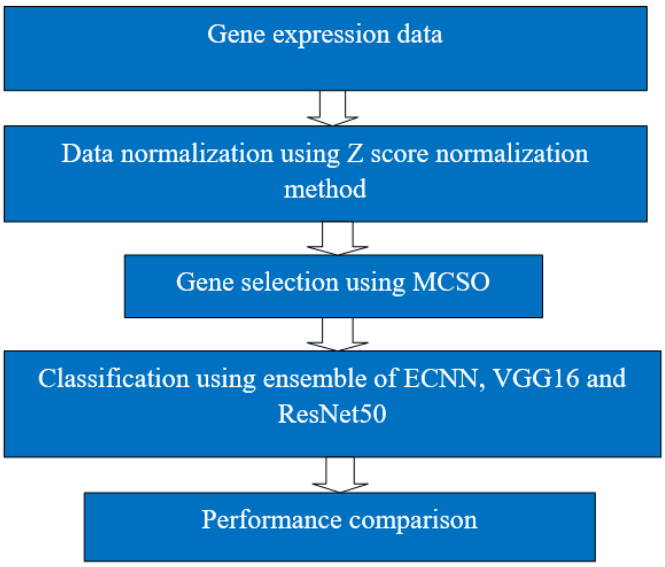

Figure 1. Over all structural design of the suggested framework

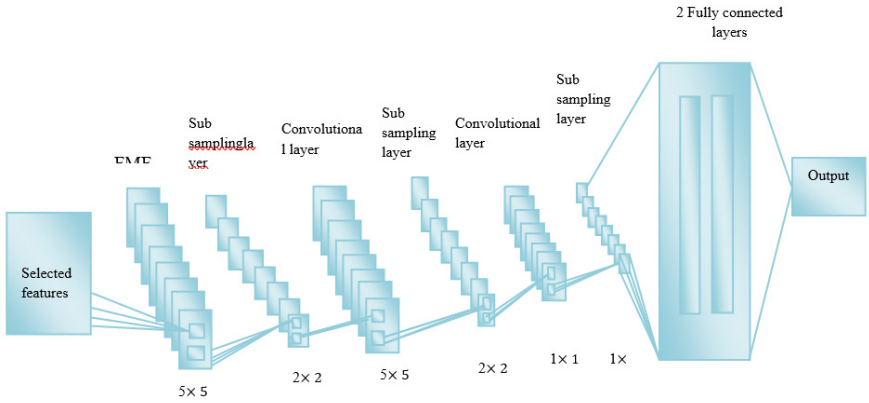

Figure.2. Convolutional Neural Network architecture

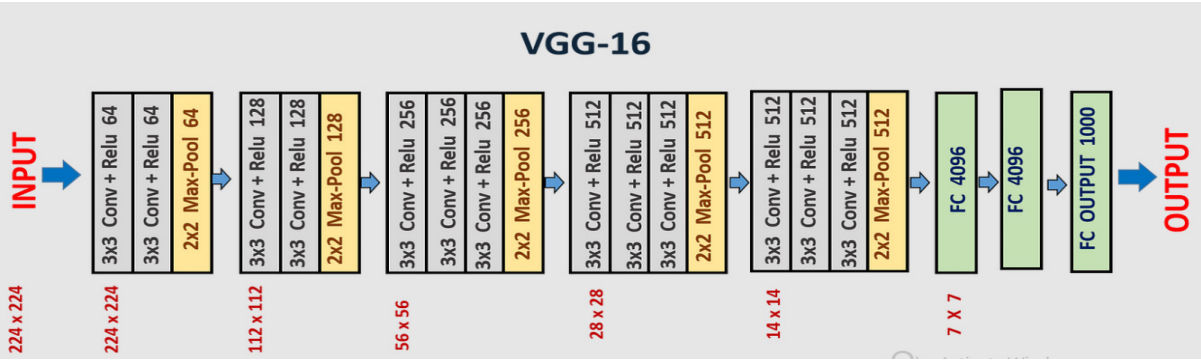

Figure 3.VGG-16 architecture

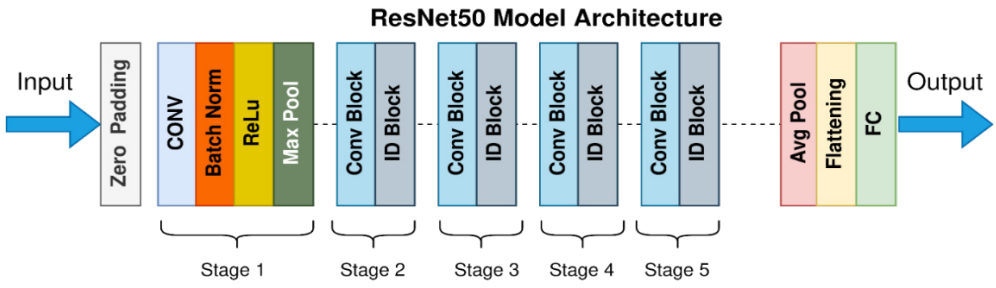

Figure 4. ResNet-50 architecture

Table.1. Performance comparison results

Metrics	Methods				
	MIMAGA	SMO	MLSTM	ICNN	EDL
Accuracy	81.22	84.46	87.12	92.78	94.84
Precision	79.89	80.85	88.78	93.06	98.19
Sensitivity	76.66	79.43	86.65	91.96	97.09
F Measure	77.14	79.01	87.34	92.51	97.63

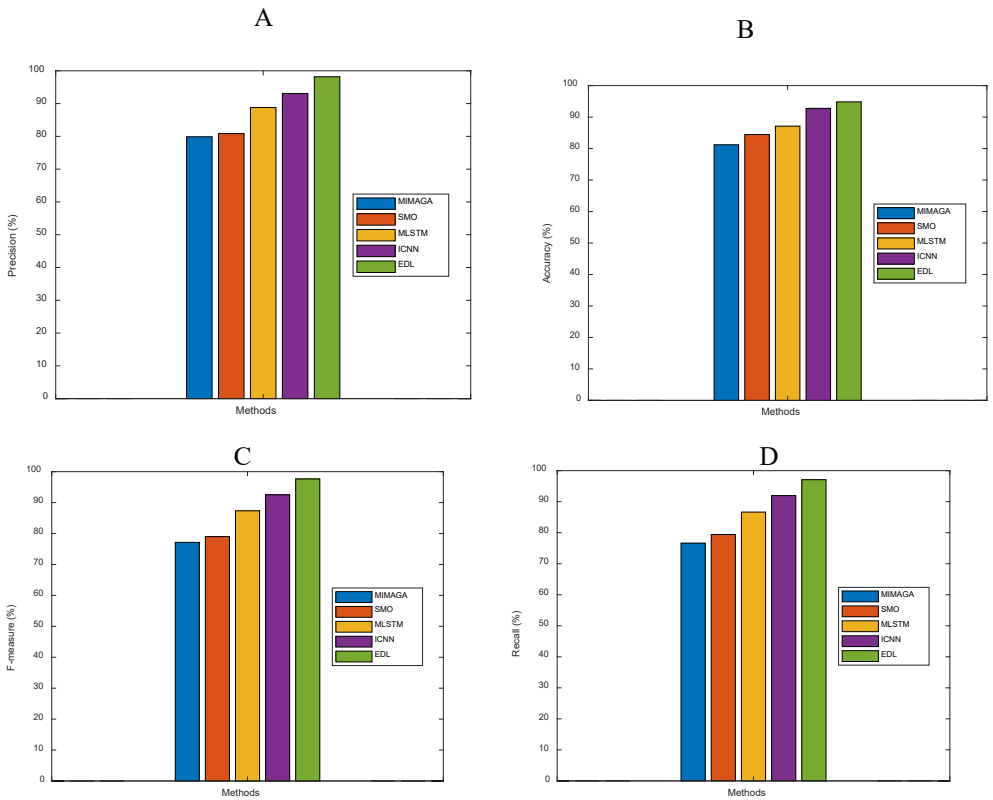

Figure 5. (A) Accuracy results, (B) Precision results, (C) Sensitivity levels, (D) F-measure results.

Results and Discussion

The rising death rates from lung cancer despite advancements in treatments highlight the need for improved diagnostic methods. Microarray analysis and gene expression profiling enable precise categorization of lung cancer, optimizing treatment. A comprehensive approach utilizing data normalization, gene selection, and ensemble models has shown promise in enhancing accuracy and reducing costs. This study provided a system for the categorization of lung cancer to address classification issues.

The study utilized Matlab 2013b to implement the proposed framework and compared the performance metrics of the Enhanced Deep Learning (EDL) system against current methodologies such as MIMAGA, SMO, MLSTM, and ICNN using the Kent Ridge Bio-Medical Dataset. This dataset comprised GE information from 10 non-neoplastic lung samples and 86 primary lung adenocarcinoma samples, featuring a total of 7129 genes. A 70-30 split was adopted, with 30% allocated as the test set and 70% as the training set for comprehensive evaluation. The outcomes of this comparative analysis are presented in Table 1.

In the assessment of machine learning (ML) and deep learning (DL) methods, accuracy stands out as a critical parameter. Five distinct approaches were applied to original input photos for evaluation. Figure 5A depicts the accuracy of various classifiers. The recommended Enhanced Deep Learning (EDL) Classifier achieved 94.84% accuracy, surpassing MIMAGA, SMO, MLSTM, and ICNN with accuracies of 81.22%, 84.46%, 87.12%, and 92.78%, respectively. This enhancement in accuracy can be attributed to two primary factors. Firstly, the suggested model employs data normalization using the z-score approach, which outperforms other standard algorithms, thus improving classification accuracy. Secondly, the utilization of a mutation operator in the Modified Chicken Swarm Optimization (MCO) contributes to enhanced accuracy.

Classifier efficiency was further evaluated using precision as an additional performance criterion. The EDL approach achieved a remarkable precision outcome exceeding 98.19% during testing (Figure 5B). In contrast, MLSTM, ICNN, SMO, and MIMAGA exhibited lower precision scores of 79.89%, 80.85%, 88.78%, and 93.06%, respectively. Notably, the ICNN classifier, incorporating a fuzzy function for weight value calculation, significantly improved precision results.

After implementing the data preprocessing approach, the suggested EDL model demonstrates superior performance over current MIMAGA, SMO, MLSTM, and ICNN systems in terms of specificity. This enhancement can be attributed to the crucial role of scale normalization in improving classifier accuracy. Sensitivity levels of the methods are depicted in Figure 5C, where the EDL strategy achieves a sensitivity rate of 97.09%, surpassing the rates of 76.66%, 79.43%, 86.65%, and 91.96% for MIMAGA, SMO, MLSTM, and ICNN systems, respectively.

The F measure, which represents the harmonic mean of precision and recall scores, serves as a comprehensive evaluation metric. Figure 5D contrasts the F measure of the suggested system with those of MIMAGA, SMO, MLSTM, and ICNN. The EDL model achieves the highest F measure at 97.63%, followed by ICNN at 92.51%, MLSTM at 87.34%, and SMO at 79.01%. The MIMAGA algorithm obtains the lowest F-measure score of 77.14%. This highlights the effectiveness of deep learning classifiers, which leverage appropriate filter sizes and weights to enhance classification accuracy and F measure results.

Conclusion

In conclusion, lung cancer remains a significant cause of mortality in China, necessitating advancements in diagnostic methods. The development of deep neural networks leveraging gene expression data offers a promising solution for early lung cancer diagnosis. This study's framework, incorporating Z-score normalization, Modified Chicken Swarm Optimization, and an ensemble of ECNN, VGG16, and ResNet50 models, achieved a notable 98.50% accuracy. Future work should address dimensionality reduction to prevent overfitting, further enhancing the efficacy of deep learning models in lung cancer classification.

Author contributions

V.Y. performed the methodology, collected and analyzed the data, and wrote the original draft. D.M. conceptualized the study, supervised the work, reviewed and edited the writing, and managed the project. V.Y. also interpreted the data and contributed to the review and editing of the writing. D.M. collected and analyzed additional data.

Acknowledgment

The authors thanked the Department.

Competing financial interests

The authors have no conflict of interest.

References

Albahar, M.A., (2019). Skin lesion classification using convolutional neural network with novel regularizer. IEEE Access, 7, pp. 38306-38313. https://doi.org/10.1109/ACCESS.2019.2906241.

Al-Haija, Q.A. and Adebanjo, A., (2020). Breast cancer diagnosis in histopathological images using ResNet-50 convolutional neural network. In IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), pp. 1-7. https://doi.org/10.1109/IEMTRONICS51293.2020.9216455.

Almugren, N. and Alshamlan, H., (2019). A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. IEEE access, 7, pp. 78533-78548. https://doi.org/10.1109/ACCESS.2019.2922987.

Arunkumar, C. and Ramakrishnan, S., (2018). Attribute selection using fuzzy roughset based customized similarity measure for lung cancer microarray gene expression data. Future Computing and Informatics Journal, 3(1), pp. 131-142. https://doi.org/10.1016/j.fcij.2018.02.002.

Azzawi, H., Hou, J., Alanni, R. and Xiang, Y., (2019). A hybrid neural network approach for lung cancer classification with gene expression dataset and prior biological knowledge. In Machine Learning for Networking: First International Conference, Revised Selected Papers 1, pp. 279-293. https://doi.org/10.1007/978-3-030-19945-6_20.

Azzawi, H., Hou, J., Alanni, R., Xiang, Y., Abdu-Aljabar, R. and Azzawi, A., (2017). Multiclass lung cancer diagnosis by gene expression programming and microarray datasets. In Advanced Data Mining and Applications: 13th International Conference, ADMA, Proceedings 13, pp. 541-553. https://doi.org/10.1007/978-3-319-69179-4_38.

Azzawi, H., Hou, J., Alnnni, R. and Xiang, Y., (2018). SBC: a new strategy for multiclass lung cancer classification based on tumour structural information and microarray data. In IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), pp. 68-73. https://doi.org/10.1109/ICIS.2018.8466448.

Cahyaningrum, K. and Astuti, W., (2020). Microarray gene expression classification for cancer detection using artificial neural networks and genetic algorithm hybrid intelligence. In international conference on data science and its applications (ICoDSA), pp. 1-7. https://doi.org/10.1109/ICoDSA50139.2020.9213051.

Chaudhari, P. and Agarwal, H., (2018). Improving feature selection using elite breeding QPSO on gene data set for cancer classification. In Intelligent Engineering Informatics: Proceedings of the 6th International Conference on FICTA, pp. 209-219. https://doi.org/10.1007/978-981-10-7566-7_22.

Dass, M.V., Rasheed, M.A. and Ali, M.M., (2014). Classification of lung cancer subtypes by data mining technique. In Proceedings of the international conference on control, instrumentation, energy and communication (CIEC), pp. 558-562. https://doi.org/10.1109/CIEC.2014.6959151.

Diaz, J.M., Pinon, R.C. and Solano, G., (2014). Lung cancer classification using genetic algorithm to optimize prediction models. In IISA, The 5th International Conference on Information, Intelligence, Systems and Applications pp. 1-6. https://doi.org/10.1109/IISA.2014.6878770.

Hafez, A.I., Zawbaa, H.M., Emary, E., Mahmoud, H.A. and Hassanien, A.E., (2015). An innovative approach for feature selection based on chicken swarm optimization. In 7th international conference of soft computing and pattern recognition (SoCPaR), pp. 19-24. https://doi.org/10.1109/SOCPAR.2015.7492775.

Haznedar, B., Arslan, M.T. and Kalinli, A., (2021). Optimizing ANFIS using simulated annealing algorithm for classification of microarray gene expression cancer data. Medical & Biological Engineering & Computing, 59, pp. 497-509. https://doi.org/10.1007/s11517-021-02331-z.

Hu, F., Zhou, Y., Wang, Q., Yang, Z., Shi, Y. and Chi, Q., (2019). Gene expression classification of lung adenocarcinoma into molecular subtypes. IEEE/ACM transactions on computational biology and bioinformatics, 17(4), pp. 1187-1197. https://doi.org/10.1109/TCBB.2019.2905553.

Jinthanasatian, P., Auephanwiriyakul, S. and Theera-Umpon, N., (2017). Microarray data classification using neuro-fuzzy classifier with firefly algorithm. In IEEE symposium series on computational intelligence (SSCI), pp. 1-6. https://doi.org/10.1109/SSCI.2017.8280967.

Kaur, T. and Gandhi, T.K., (2019). Automated brain image classification based on VGG-16 and transfer learning. In international conference on information technology (ICIT), pp. 94-98. https://doi.org/10.1109/ICIT48102.2019.00023.

Metwalli, A.S., Shen, W. and Wu, C.Q., (2020). Food image recognition based on densely connected convolutional neural networks. In international conference on artificial intelligence in information and communication (ICAIIC), pp. 027-032. https://doi.org/10.1109/ICAIIC48513.2020.9065281.

Moldovan, D., (2020). Cervical cancer diagnosis using a chicken swarm optimization based machine learning method. In international conference on e-health and bioengineering (EHB), pp. 1-4. https://doi.org/10.1109/EHB50910.2020.9280215.

Salem, H., Attiya, G. and El-Fishawy, N., (2017). Classification of human cancer diseases by gene expression profiles. Applied Soft Computing, 50, pp. 124-134. https://doi.org/10.1016/j.asoc.2016.11.026.

Tian, X. and Chen, C., (2019). Modulation pattern recognition based on Resnet50 neural network. In IEEE 2nd International Conference on Information Communication and Signal Processing (ICICSP), pp. 34-38. https://doi.org/10.1109/ICICSP48821.2019.8958555.

Tripathi, A.K., Garg, P., Tripathy, A., Vats, N., Gupta, D. and Khanna, A., (2020). Prediction of cervical cancer using chicken swarm optimization. In International Conference on Innovative Computing and Communications: Proceedings of ICICC, 1, pp. 591-604. https://doi.org/10.1007/978-981-15-1286-5_51.

Venkatesan, C., Balamurugan, D., Thamaraimanalan, T. and Ramkumar, M., (2022). Efficient machine learning technique for tumor classification based on gene expression data. In 8th International Conference on Advanced Computing and Communication Systems (ICACCS), 1, pp. 1982-1986. https://doi.org/10.1109/ICACCS54159.2022.9785294.

Wang, Y., Li, X. and Ruiz, R., (2018). Weighted general group lasso for gene selection in cancer classification. IEEE transactions on cybernetics, 49(8), pp. 2860-2873. https://doi.org/10.1109/TCYB.2018.2829811.

Yuan, F., Lu, L. and 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), pp. 165822. https://doi.org/10.1016/j.bbadis.2020.165822.

Zarlis, M., Yanto, I.T.R. and Hartama, D., (2016). A framework of training ANFIS using chicken swarm optimization for solving classification problems. In International conference on informatics and computing (ICIC), pp. 437-441. https://doi.org/10.1109/IAC.2016.7905759.