A Hybrid Deep Learning Based Approach for Lung Cancer Classification by using Microarray Data Analysis

Bhanumathi S
Research Scholar, Department of Computer Science & Engineering, Assistant Professor
Department of Information Science & Engg
SJCIT, Chickballapur
Visvesvaraya Technological University,
Belagavi, Karnataka Email: snc.boe.cse@gmail.com
Email:sbhanureddy14@gmail.com

Dr.S N Chandrashekara,
Professor & Head
Department of Computer Science & Engg
C B IT, Kolar-563101

Abstract

Lung cancer is one of the serious causes of the mortality in cancer patients. Because of the fast improvement of DNA microarray technology, researchers are able to evaluateextensively huge gene expression datasetin one experiment.However, because gene expression databases sometimes contain huge numbers of genes in a small number of tissues, categorising microarray data for cancer detection and prevention is a big challenge.In order to build a robust gene signature from microarray data, several researchers have examined a variety of gene selection methodologies for the prediction of cancer recurrence. However, the accuracy of gene classification still remains a challenging task. Thus, here we introduce a novel hybrid technique for microarray data classification. This approach uses a feature selection, feature ranking and deep learning scheme to learn the gene attributes and predict the lung cancer.

Keywords: microarray data, lung cancer, deep learning, feature ranking and selection

DOI Number: 10.14704/nq.2022.20.11.NQ66044 **NeuroQuantology 2022; 20(11): 414-424**

1. Introduction

Nowadays, we are noticing several health related issues where some issues are curable and normal whereas some health concerns are life-threatening and require best treatment to overcome those issues. In this line, cancer is considered as serious public health issue.Lung cancer disease is a kind of primary bronchogenic carcinoma which is classified into two categories: small cell carcinoma and non-small cell carcinoma. The small cell carcinoma accounts 10%-15% and NSCLC accounts 85-90% of total cancer cases [1]. The NSCLC type of cancer include two sub-types i.e., adenocarcinoma and squamous cell carcinoma.Cancer is a critical and multi-faceted disease by virtue which is categorized by the uncontrolled and progressive nature of cell divisions. Cancer is among the most common causes of mortality in the world. A case study presented in [2] has reported 8.2 million deaths and 15 million cases each year.

Currently, the technological advancements have helped diagnosing lung cancer in infancy stages.The rate of lung cancer is 15% and around 75% patients are successfully diagnosed who are suffered from cancer at metastatic stage [3]. However, the poor prediction performance is obtained for the patients who suffered from advanced stage lung cancer. Moreover, the 5-year survival rate for a novel specialised anti-tumour medication that works as a checkpoint inhibitor was likewise less than 15% [4]. Several factors are present which affects the severity and cause the lung cancer are like radon exposure, genetic mutations, tobacco consumption, unbalanced diet, smoking, tobacco and many more. Several techniques are present to treat the early-

eISSN 1303-5150

stage tumours by adopting chemotherapy, targeted therapy, chemotherapy-immunotherapy.

Current technology is focusing on the early prediction of cancers in early stage which includes computer vision and data mining based approaches. Computer vision based approaches require data in image or video form such as chest X-Ray screening, Low-dose CT etc. where various techniques are used for extracting the significant features and matching those features by utilizingartificial intelligence techniques such as Deep Learning (DL) technique applied on CT images [4], image segmentation and morphological feature extraction based approach by using deep learning scheme [5] and many more such as mentioned in [6]. Similarly, the data mining based techniques are also widely adopted for classification of the lung cancer on the basis of the electronic health records. Several data mining schemes are presented such as back propagation neural network [7], boosted neural network ensemble classification which uses Weight Optimized Neural Network with Maximum Likelihood Boosting (WONN-MLB) and many more [10, 11]. Nowadays, the gene expression profiling is considered as a promising technique in the field of biomedical data processing. The gene expression profiling comes under the subject of molecular biology which provides the measurement of 1000s of genes at single shot to obtain a global representation of cellular functions of specific cell. Moreover, gene profiling shows how cells react to the particular treatment. The technique for detection of the expression of thousands of genesin one shot is known as microarray data. Nowadays, microarray data based gene profiling is considered as a novel technique to obtain the gene expression profiling. The advancements in microbiology the microarray technology provides the information of DNA, RNA and protein to detect the various types of tumours in early stages. The efficient processing of these techniques helps inreducing the mortality rate of lung cancer patients. The microarray technology plays important role to determine the activity stages of genes as active, hyperactive or inactive in various tissues. These genes are classified into two or more groups and the information of DNA, RNA, and protein is obtained to detect the formation of tumour in its earlier stages. Several techniques have been presented for lung cancer detection such as Ghosh et al. [13] developed a meta-heuristic approach for feature selection where ensemble feature selection and genetic algorithm based optimization schemes are used to classify the obtained attribute patterns. Finally, these features are classified with the help of machine learning algorithm. In [23] authors presented a fuzzy logic based attribute selection method by using fuzzy logic to measure the similarity between microarray gene expression data. Alanni et al. [24] presented a novel approach for gene selection for cancer classification by using microarray datasets. Mehmood et al. [25] presented a clustering algorithm with the help of fast searching algorithm which merges the local density features for gene expression microarray dataset. The existing schemes suffer from various challenging issues;therefore, we present a novel approach that is based on deep learning technique for lung cancer detection by using microarray data analysis. The main contributions of this approach are following:

- We presented a concise literature review of existing schemes which are based on the microarray data analysis and identified the drawbacks
- a new approach for attribute selection for dimensionality reduction is introduced
- Finally, a deep learningbased approach is presented to detect the lung cancer.

eISSN 1303-5150

Rest of the article is organized in following order: section II reviews the literature about present and prevailingapproaches, section III presents the proposed solution to detect the lung cancer by usingDL technique, section IV presents the experimental study and obtained performance is compared with the existing schemes, finally, section V presents the concluding remarks and future scope for lung cancer detection by using microarray data.

1. Literature survey

In this section, we present the description about existing methods of lung cancer detection and classification using MLmethods. Also, we study about the microarray data based gene profiling schemes to analyse the genes to detect the lung cancer.

Sayed et al. [12] considered high dimensional microarray dataset and focused on feature selection for cancer detection. The main challenge of microarray data processing occurs due to its high dimensionality and more number of attributes. Authors presented an ensemble feature selection approach to overcome these dimensionality issues. This ensemble approach is an integration of the$t-$ test and Genetic Algorithm (GA). The t-test is used for data pre-processing and nested GA is utilized to obtain the optimal feature subset where it combines the data from two different datasets. The nested GA contains two different types of datasets. The outer GAoperates on gene expression dataset and inner genetic algorithm focuses on DNA Methylation datasets. Moreover, it applies incremental feature selection strategy.Ghosh et al. [13] also reported that the high dimensionality of these dataset is a challenging issue because this type of dataset contains irrelevant and redundant attributes which increases the data processing complexity. The feature selection scheme is considered as a promising solution for these issues which is an NP-Hard problem. The NP-hard problem can be resolved by using meta-heuristic search approaches thus authors developed a two-stage feature selection strategy in microarray dataset. In first phase, it uses an ensemble of filtering scheme where it considers union and intersection of top-n attributes based on ReliefF, chi-square and symmetrical uncertainty. This ensemble helps to aggregate the ranking outcome. Later, a genetic algorithm is applied on these intersections and union points to fine-tune the results. Lai et al. [14] studied about non-small cell lung cancer and suggested that accurate prediction of lung cancer has become an important task in biomedical field which helps to design the therapeutic strategies to diagnose the cancer. Thus, authors developed a deep learning scheme based on the Deep Neural Network (DNN) to analyze the gene expression data. The DNN scheme considers the gene biomarkers to develop the DNN based survival prediction. Azzawi et al. [15] focused on prediction of lung cancer disease by using microarray data analysis by using gene expression programming based model. In this scheme, authors presented two schemes for gene selection to obtain the significant attributes corresponding to the different Gene Expression Programming (GEP) prediction model where SVM, Multi-Layer Perceptron (MLP) and Radial Basis Function (RBF)neural network learning schemes are adopted.

Aydadenta et al. [16] discussed that microarray data is helpful to analyze thousands of gene samples simultaneously however the dimensionality of these type of dataset is a challenging issue. Thus, authors developed a feature selection scheme by selecting the high dimensional features which are having higher correlation with other features. This scheme uses k-means clustering approach which minimizes the redundancy by grouping the attributes in a single cluster. Further, the Relief technique is utilized to obtain the greatest scoring cluster. Further, the Random forest classification algorithm is

eISSN 1303-5150

applied. Aydadentan et al. [17]utilized data mining techniques likeSVM, ANN Naïve Bayes (NB), kNN and C4.5. These technique shows that the efficient machine learning algorithm can help to improvise the classification results. AbdElNabi et al. [18] concentrated on early detection of cancer by developing a smart systemto diagnose cancer in preliminary stageon the basis of gene expression profiles which are retrieved using DNA microarrays. In addition, the authors used information gain to extract relevant characteristics from the input dataset. Following that, the grey wolf optimization approach is used to minimise these traits. Lastly, the SVM classifier is applied to classify the cancer.

Wang et al. [19] recommended usage of supervised learning scheme for detecting the lung cancer to decrease the death rate. Authors developed a random forest with self-paced learning bootstrap algorithm. Specifically, this scheme uses random forest classification scheme and later high-to-low quality samples are embedded using self-paced learning. Mallick et al. [20] discussed the advantages of optimization schemes when combined with machine learning algorithm resulting in improving the classification performance. To accomplish this, authors presented a combined scheme of artificial neural network and ant colony optimization. Here, the ant colony optimization helps to fine-tune the parameters of ANN. Further, principal component analysis (PCA) is applied for dimension reduction. The obtained dimensionality reduced dataset is further trained using Functional Link Artificial Neural Network (FLANN).

Purbolaksono et al. [21] developed a new approach for feature selection where it doesn't discard every information from its attributes. Thus it preserves the finer information precisely. For this purpose, it uses mutual information to reduce the dimensions. Further, Bayesian theorem based approach which uses statistical and probability approach. This combination improves the classification performance. Zadeh et al. [22] focused on Triple-negative breast cancer (TNBC) and tried to predict the Basal-like breast cancers (BLBCs) and non- Basal-like breast cancers (BLBCs) by applying machine learning classification strategy. The data processing scheme includes data cleaning, dimension reduction, and variable selection. The feature selection scheme uses chi-square, decision tree, LASSO and principal component analysis. Finally, the obtained dataset is processed through several classification schemes to classify the attributes and classification performance is measured.

2. Proposed Model

This section presents the proposed solution for lung cancer detection by using deep learning based data mining strategies. In first phase, we present a novel feature selection approach and the optimal features are ranked by applying feature ranking method. Later, these attributes are processed through the deep learning based classification model for n-fold cross validation to find the best solution for classification.

2.1. Feature Selection

The gene microarray data is huge in dimension where various attributes which doesn't have the significant impact on the performance of microarray data analysis. The existing schemes have reported that discarding the attributes improves the classification accuracy and minimizes the computation time. moreover, this helps to scale the classification algorithm when we are adding more number of samples to existing dataset.

Let us consider that we have a matrix as $X^{m\times n} = \{x_{i,j}\}$ which contains total m number of attributes and n number of samples. These samples are obtained from different groups which are denoted by a target class as $X^{m\times n} = \left[X_1^{m\times n_1} X_2^{m\times n_2}, \dots, X_p^{m\times n_p}\right]$ where each matrix $X_i^{m\times n_i}$ contains various samples

eISSN 1303-5150

from same group and $n_1 + n_2 + .. , + n_p = n$. Here selection of attributes from these groups is tedious task. We focus on selection of most significant features subset as $S^{k \times n} \in X^{m \times n}$, $k \ll m$. In order to obtain the most effective features, we present a novel feature selection based technique which is considers the uncertainties of attributes and incorporates ReliefF algorithm to estimate the attributes. The main aim of incorporating gene uncertainty information is to improve the scalability of proposed approach, moreover, the attributes are obtained from different sources thus they are diverse in nature. Furthermore, we incorporate a filtering based method which has three novel criterions such as the existing filtering based methods do not consider the redundancy between selected genes whereas this method adopts the redundancy and considers the redundancy which minimizes the scenario of discarding the effective attributes. The traditional feature selection algorithms are based on the mutual information but these techniques are biased towards the gene with larger values thus selected features are not efficient and these types of schemes requires additional attributes to improve the performance. This issue is resolved in proposed approach because it defines a range of mutual information to select the optimal features. To overcome these issues, we present a new technique for feature extraction.

The mutual information is considered as the information which is shared by two variables which helps to identify the similarity between considered variables. The mutual information can be expressed as:

$$I(X,Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (1$$

Where $p(x,y)$ denotes the joint probability distribution function for variable x and y, similarly , the $p(x)$ is the edge probability distribution of X, $p(Y)$ is the probability distribution of Y. The mutual information can be expressed as

$$I(X,Y)H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y) \quad (2$$

Where $H(X|Y)$ denotes the conditional probability, $H(X)$ and $H(Y)$ represent the edge entropy and $H(X,Y)$ represents the joint entropy of the input data. With the help of this, the range of mutual information is defined as:

$$0 \leq I(f_i, f_s) \leq \min\{H(f_i), H(f_s)\} \quad (3$$

As discussed before, the mutual information has the issue of biasing towards the attributes which are having higher feature values. In order to deal with this issue, we present a threshold model which is expressed as:

$$T(X,Y) = \frac{2I(X,Y)}{H(X) + H(Y)} \quad (4$$

The value of this threshold varies in an interval of [0,1]. If the threshold ratio is obtained as 1 then it denotes one random variable X and Y are completely independent and can be predicted. Similarly, if the value of threshold is 0 then it is concluded that and X and Y are independent. This process minimizes the biasing issues.

In order to perform the feature selection, we apply ReliefF which is an attributes estimation algorithm. The conventional relief algorithm considers only two-class classification problem whereas this approach is able to handle multiple classes. Mainly this scheme estimates the weights of attributes and evaluate these attribute weights with the help of multiple number of iterations. During this process, instance R is selected randomly from the dataset and it searches for k-nearest neighbour which are known as H from instances and similarly, it searches M nearest neighbours. In next stage we consider the distance parameter between these nearest neighbour points. According to the distance analysis, if the distance between R and H is less than the distance between sample R and M then this scenario is used to distinguish the instances from different classes otherwise

eISSN 1303-5150

these samples are not considered of significant use and can be discarded for further use. This process is repeated for all attributes until the weights of final feature are estimated. The weight updating can be computed as:

$$W[i] = W[i] - \sum_{j=1}^{k} \frac{diff(i,R,H_j)}{m.k} + \sum_{C \neq class\ (R)} \frac{P(C)}{1-P(class(R))} . \frac{\sum_{j=1}^{k} diff\ (i,}{mk}$$

Here, $diff$ denotes the difference of attributes between sample R and H which is computed as:

$diff(i,R,H) = \frac{value\ (i,I_1) - value\ (i,I_2)}{\max(i) - \min(i)}$, $P(C)$ is the probability of class C and $1 - P(class(R))$ denotes the sum probability.

Further, we apply genetic algorithm based model for dimension reduction to minimize the dimensionality of the gene dataset. Genetic algorithm is a widely adopted population based stochastic optimization technique. This approach is widely adopted in various applications. however, this scheme suffers from two main drawbacks which are slow convergence and achieving the optimal points.

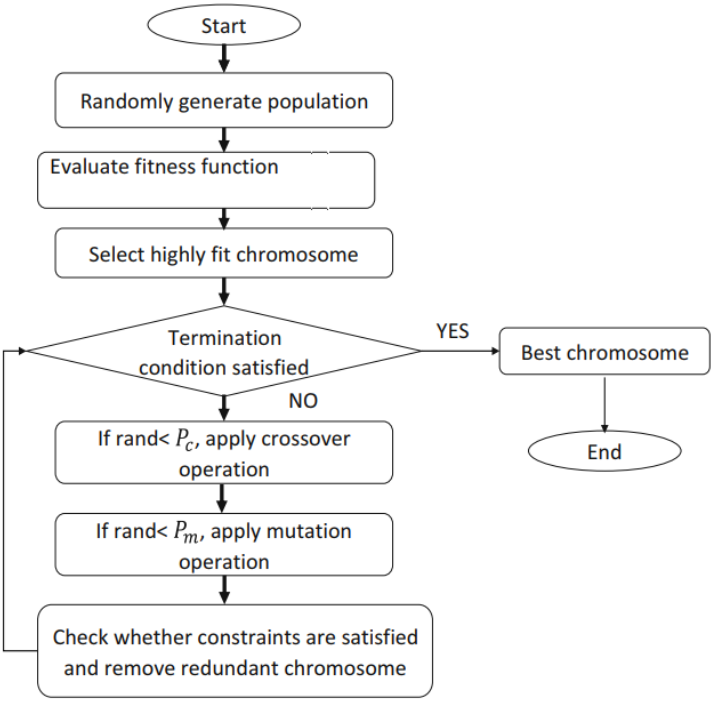

Fig.1. genetic algorithm

In this work, we have adopted genetic algorithm from [26] which contains new crossover and mutation operations to mitigate the existing challenges. According to genetic algorithm the outcome of solution depends on the probability of crossover (P_c) and probability of mutation (P_m) which affects the accuracy and convergence speed of the genetic algorithm. The technique mentioned in [26] updates the values of P_m and P_c dynamically with the help of adaptive process. The probability of crossover can be computed as:

$$P_c = \begin{cases} \left[k_1 . \left(\frac{f_{max} - f'}{f_{max} - f_{avg}}\right)\right] - k5, f' \geq f\_avg \\ k_2, f' < f_{avg} \end{cases} \quad (6)$$

eISSN 1303-5150

Similarly, the probability of mutation can be computed as:

$$P_m = \begin{cases} \left[k_3.\left(\frac{f_{max} - f}{f_{max} - f_{avg}}\right)\right] - k6, f' \geq f\_avg \\ k_4, \quad f < f_{avg} \end{cases} \quad (7)$$

where f_{max} represents the maximum fitness value, f_{avg} is the average fitness value, f' is the average fitness of three parents which are considered during selection operation. k_1 to k_6 represents the constant values which are ranging between 0 and 1. Below given figure 1 depicts the overall working of genetic algorithm. According to this process, first of all, we generate the initial population and evaluate their fitness. The population which is having high fitness value are selected for further process. Later, crossover and mutation operations are applied and fitness is measured. This process is repeated until the best fit conditions are satisfied.

This process generates some attributes which are further processed through the threshold based weight evaluation method. This combination of weight factor is defined as:

$$R_{i,c} = \mu T_{i,c} + (1 - \mu) W_i \quad (8)$$

Where $\mu \in [0,1]$. The above given expression generates the sorted attributes which are processed through the deep learning model for classification.

2.2. Deep Learning (DL) Classification

In this section we describe the DLframework used for classification. The proposed architecture is divided into two branches of deep learning modules where one module utilizes fully connected NN and second module uses CNN. The fully connected NN consist of three fully connected layers where each layer has 100 neurons. These layers use softmax activation function. Lastly, a layer with 11 neurons is generated with softmax activation function which generates the probability of cancer. Below given figure 1 depicts the architecture frameworkutilized in this approach.

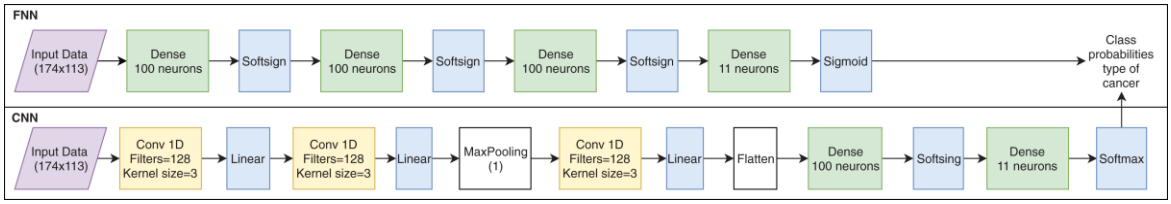

Fig.2. deep learning framework used for classification

3. Results and Discussion

Here, we present the experimental study where we measured the performance of proposed approach and compared the obtained outcome of proposed approach with the existing schemes. The performance is measured for different types of biological datasets.

Second, as a fitness function for the suggested method, we choose the best classification technique. Finally, we looked at two classification algorithms, SVM and Naive Bayes, in this regard. When the outcomes of the two classifiers are compared, it is clear that the SVM classifier surpasses the NB classifier. The reason being that the features in the NB classifier are considered dependent, which reduces the classifier's efficiency. Lastly, the suggested technique's findings are compared to those of previously published gene selection techniques. The classifying performance is measured using two classifiers, SVM and NB, which act as fitness functions in the MPAGA technique. On 10 gene datasets, the performance is measured using four parameters: accuracy, sensitivity, specificity, and F-measure. Eqs. (9) to (11) define these performance measurements (12):

	Positive	Negative	Total
Positive	T_p	F_p	$T_p + F_p$
Negative	F_N	T_N	$F_N + T_N$
Total	$T_P + F_N$	$T_P + T_N$	

eISSN 1303-5150

The detection accuracy is calculatedon the basic of values obtained (as mentioned in confusion matrix). The accuracy can be computed as:

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \quad (9)$$

Similarly, we compute the sensitivity performance by using confusion matrix. The sensitivity can be expressed as:

$$Sensitivity = \frac{T_P}{T_P + F_N} \quad (10)$$

The specificity can be computed as:

$$Specificity = \frac{T_N}{T_N + F_P} \quad (11)$$

And, F-measure is can be computed as

$$F - measure = \frac{2 \times T_P}{2 \times T_P + F_N + F_P} \quad (12)$$

3.1. Dataset details

To measure the efficiency of projectedscheme, we chosen microarray dataset as Breast cancer, Colon cancer, DLBCL, Leukaemia, Tumor_1, and Lung cancer. These dataset contains the large number of genes as attributes. Table 1 provides the description of these dataset where total instances, number of genes as attributes, and total classes.

Dataset	***Instances***	***Genes***	***Classes***	***Dataset***	***Instances***	***Genes***	***Classes***
Breast cancer	97	24,481	2	*SBRCT*	83	2308	4
Colon cancer	62	200	2	*Lung cancer*	203	12,600	5
DLBCL	77	5469	2	*Brain Tumor_1*	90	5920	5
Leukaemia	72	7129	2	*9_Tumors*	60	5726	9
Prostate tumor	102	10,509	2	*11_Tumors*	174	15,533	11

Based on these dataset, we conducted experiments and measured the average performance for overall scenario in terms of accuracy, sensitivity, specificity and f-measure. The obtained performance is compared with the existing schemes such as support vector machine and Naïve Bayes classifier as mentioned in [26]. The comparative performance is presented in below given table 2.

Table.2. average comparative performance in terms of classification accuracy

Dataset	*Naïve Bayes*				*SVM*				*Proposed Approach*			
	Acc	*Sn*	*Sp*	*F − Measure*	*Acc*	*Sn*	*Sp*	*F − Measure*	*Acc*	*Sn*	*Sp*	*F − measure*
Breast cancer	90.46	91.25	89.04	88.12	90.98	90.01	91.42	90.62	95.23	94.11	91.15	93.20
Colon cancer	92.17	93.54	92.64	94.65	93.78	93.97	94.62	93.97	94.51	93.25	90.85	92.50
DLBCL	91.99	91.74	93.57	92.17	94.97	95.01	95.88	93.01	96.25	95.50	96.50	91.50
SBRCT	90.02	55.54	87.33	86.77	97.57	90.52	90.22	90.94	92.50	96.20	94.10	96.50
Lung cancer	93.47	94.05	94.68	95.87	94.52	96.22	96.57	96.01	95.60	95.80	95.50	92.20
Leukaemia	94.16	89.42	88.39	86.87	93.44	91.47	92.87	92.75	93.50	95.60	96.50	96.50
Brain_tumor_1	84.32	84.68	84.90	85.43	86.87	86.01	87.98	89.54	91.85	96.20	95.15	95.50
11_tumor	87.98	87.12	87.54	88.76	89.01	88.34	89.34	88.54	92.50	96.1	96.20	94.30
9_tumor	75.98	75.01	76.94	77.86	76.34	77.45	78.98	76.12	90.10	95.25	95.10	95.50
Prostate cancer	92.54	91.08	92.65	90.87	93.76	93.89	93.65	94.01	94.55	93.50	94.20	96.50

In this work, we have used a dimension reduction and feature selection scheme based on which we select the top 50 attributes as genes to measure the performance. The obtained performance is presented in below given table 3.

eISSN 1303-5150

Table.3. comparative performance for top 50 genes

Classifier	*Feature selection*	*Breast*	*Colon*	*DLBCL*	*SBRCT*	*Lung*	*Leukaemia*	*Brain_tumor_1*	*11_tumor*	*9_tumor*	*Prostate cancer*
SVM	*CMIM*	81.01	80.42	92.54	92.99	90.52	90.67	83.65	84.36	68.95	88.64
	JMI	70.54	73.62	78.66	90.54	91.67	91.53	81.36	88.95	64.98	87.34
	mRMR	83.23	81.63	88.52	93.48	91.36	92.52	83.65	80.11	69.85	86.85
	DISR	76.52	79.33	80.01	92.37	91.55	90.94	81.65	81.96	70.90	81.97
	Relief − F	73.99	82.57	80.67	83.54	85.66	80.74	84.03	82.36	71.35	88.16
	[26]	91.08	86.99	93.06	96.08	92.14	93.57	84.78	83.95	72.18	85.17
NB	*CMIM*	79.01	70.25	81.35	80.44	76.52	82.99	81.03	83.95	64.94	87.36
	JMI	81.54	73.78	81.99	82.74	75.05	81.04	80.65	84.69	65.85	86.95
	mRMR	87.07	81.46	82.52	85.74	86.34	91.37	81.56	84.36	64.96	82.74
	DISR	84.52	79.19	82.01	81.54	83.22	90.09	80.36	78.99	68.32	79.65
	Relief − F	82.54	83.11	90.99	91.54	89.67	92.98	82.54	79.26	70.15	84.26
	[26]	89.52	84.85	92.94	92.89	91.52	93.02	81.92	79.38	71.36	83.74
Deep Learning	*Proposed*	97.23	95.22	97.41	93.50	94.23	95.11	91.08	89.55	85.63	94.20

Finally, we present a comparative analysis by considering existing optimization based feature selection algorithm where we have considered MPAGA, GA and adaptive genetic algorithm [26]. Below given table 4 shows the comparative analysis for this scenario. In this scenario, we have measured the best, average and minimum accuracy obtained by various classification schemes.

Table.4. comparative analysis in terms of accuracy

Dataset	*Adaptive Genetic*			*GA*			*MPAGA*			*Proposed*		
	Best	*Avg.*	*Min*	*Best*	*Avg.*	*Min*	*Best*	*Avg.*	*Min*	*Best*	*Avg.*	*Min*
Breast Cancer	99.64	94.15	87.23	90.32	87.35	86.32	86.36	84.36	81.36	99.70	89.23	87.15
Colon cancer	*99.15*	*98.87*	*93.64*	*96.51*	*95.68*	*90.99*	*93.21*	*92.47*	*90.67*	*99.421*	*99.114*	*95.52*
DLBCL	*100*	*99.51*	*97.89*	*98.42*	*97.26*	*94.76*	*95.37*	*94.36*	*93.89*	*100*	*99.55*	*98.25*
SBRCT	*99.87*	*98.93*	*96.20*	*96.03*	*94.36*	*91.27*	*91.95*	*90.65*	*88.84*	*99.89*	*99.10*	*98.51*
Leukaemia	*99.32*	*98.84*	*97.23*	*92.36*	*91.35*	*90.87*	*90.15*	*88.35*	*84.69*	*99.56*	*99.56*	*98.22*
Lung cancer	*100*	*99.52*	*97.05*	*93.85*	*90.25*	*88.46*	*88.36*	*87.96*	*85.39*	*100*	*99.68*	*98.5*
Brain Tumor	*98.24*	*96.92*	*88.41*	*89.35*	*88.36*	*84.56*	*86.35*	*84.65*	*81.17*	*98.89*	*97.59*	*94.25*
11 Tumor	*94.75*	*93.04*	*89.06*	*91.25*	*90.36*	*89.65*	*87.95*	*84.20*	*82.18*	*97.56*	*96.12*	*92.35*
9 tumor	*88.01*	*82.88*	*71.36*	*76.54*	*72.36*	*71.23*	*70.32*	*68.35*	*64.70*	*94.23*	*91.22*	*89.50*
Prostate	*99.04*	*98.42*	*91.91*	*93.88*	*90.02*	*86.35*	*88.01*	*87.96*	*79.36*	*9917*	*99.25*	*94.71*

The above comparative analysis shows that the proposed approach achieves better performance when compared with existing techniques which are based on the optimization strategies. According to feature selection strategy we have obtained the 10 attributes which are having more impact which are 33069_f_at, 39272_g_at, 33892_at, 34076_at, 31431_at, ,38489_at, 33712_at, and AFFX-BioDn-3_at.

4. Conclusion

Current the biomedical field has grown drastically and increasing advancements in technology has led to develop a high-throughput microarray data technology which is considered as a breakthrough in this field. The microarray technology helps to generate the large genomics datasets which are characterized by the number of genes. However, the high dimensionality of these datasets becomes a challenging issue to process and detect the cancer. Thus, feature selection is believedto be a capable scheme thataidsin grouping the biomarkers based on their attributes. Several techniques have been presented for feature selection but obtaining

eISSN 1303-5150

significant features still remains a tedious task. Here, we present a novel approach for dimension reduction and feature selection for microarray dataset to detect the lung cancer. The proposed approach uses feature selection, genetic algorithm based feature optimization and deep learning based classification model. The experiments performed in this research proves the efficacy of the proposed scheme while achieving better classification accuracy while evaluating against existent schemes.

References

1. Herbst, R. S., Morgensztern, D., & Boshoff, C. (2018). The biology and management of non-small cell lung cancer. Nature, 553(7689), 446-454.
2. Cancer research, U. Annual report. 2017-2018; Available from: https://www.cancerresearchuk.org/sites/default/files/cruk_annual_report_2017_18_final.pdf
3. Jacobsen, M. M., Silverstein, S. C., Quinn, M., Waterston, L. B., Thomas, C. A., Benneyan, J. C., & Han, P. K. (2017). Timeliness of access to lung cancer diagnosis and treatment: a scoping literature review. *Lung cancer*, *112*, 156-164.
4. Bai, R., Lv, Z., Xu, D., & Cui, J. (2020). Predictive biomarkers for cancer immunotherapy with immune checkpoint inhibitors. *Biomarker Research*, *8*(1), 1-17.
5. Shakeel, P. M., Burhanuddin, M. A., & Desa, M. I. (2020). Automatic lung cancer detection from CT image using improved deep neural network and ensemble classifier. *Neural Computing and Applications*, 1-14.
6. Bansal, G., Chamola, V., Narang, P., Kumar, S., & Raman, S. (2020). Deep3DSCan: Deep residual network and morphological descriptor based framework for lung cancer classification and 3D segmentation. *IET Image Processing*, *14*(7), 1240-1247.
7. Saba, T. (2020). Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges. *Journal of Infection and Public Health*, *13*(9), 1274-1289.
8. Varadharajan, R., Priyan, M. K., Panchatcharam, P., Vivekanandan, S., & Gunasekaran, M. (2018). A new approach for prediction of lung carcinoma using back propogation neural network with decision tree classifiers. Journal of Ambient Intelligence and Humanized Computing, 1-12.
9. ALzubi, J. A., Bharathikannan, B., Tanwar, S., Manikandan, R., Khanna, A., & Thaventhiran, C. (2019). Boosted neural network ensemble classification for lung cancer disease diagnosis. *Applied Soft Computing*, *80*, 579-591.
10. Ray, S. (2021). A Survey on Application of Machine Learning Algorithms in Cancer Prediction and Prognosis. In *Data Management, Analytics and Innovation* (pp. 349-361). Springer, Singapore.
11. SN, D. (2020). Impute, Select, Decision Tree and Naïve Bayes (ISE-DNC): An Ensemble Learning Approach to Classify the Lung Cancer.
12. Sayed, S., Nassef, M., Badr, A., & Farag, I. (2019). A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets. Expert Systems with Applications, 121, 233-243.
13. Ghosh, M., Adhikary, S., Ghosh, K. K., Sardar, A., Begum, S., & Sarkar, R. (2019). Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods. Medical & biological

eISSN 1303-5150

engineering & computing, 57(1), 159-176.

14. Lai, Y. H., Chen, W. N., Hsu, T. C., Lin, C., Tsao, Y., & Wu, S. (2020). Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. *Scientific reports*, *10*(1), 1-11.
15. Azzawi, H., Hou, J., Xiang, Y., & Alanni, R. (2016). Lung cancer prediction from microarray data by gene expression programming. *IET systems biology*, *10*(5), 168-178.
16. Aydadenta, H., & Adiwijaya, A. (2018). A clustering approach for feature selection in microarray data classification using random forest. *Journal of Information Processing Systems*, *14*(5), 1167-1175.
17. Aydadenta, H. (2018, March). On the classification techniques in data mining for microarray data classification. In *Journal of Physics: Conference Series* (Vol. 971, No. 1, p. 012004). IOP Publishing.
18. AbdElNabi, M. L. R., Wajeeh Jasim, M., M EL-Bakry, H., Taha, H. N., & M Khalifa, N. E. (2020). Breast and colon cancer classification from gene expression profiles using data mining techniques. *Symmetry*, *12*(3), 408.
19. Wang, Q., Zhou, Y., Ding, W., Zhang, Z., Muhammad, K., & Cao, Z. (2020). Random forest with self-paced bootstrap learning in lung cancer prognosis. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *16*(1s), 1-12.
20. Mallick, P. K., Satapathy, S. K., Mishra, S., Panda, A. R., & Mishra, D. (2021). Feature selection and classification for microarray data using ACO-FLANN framework. In *Intelligent and cloud computing* (pp. 491-501). Springer, Singapore.
21. Purbolaksono, M. D., Widiastuti, K. C., Mubarok, M. S., & Ma'ruf, F. A. (2018, March). Implementation of mutual information and bayes theorem for classification microarray data. In *Journal of Physics: Conference Series* (Vol. 971, No. 1, p. 012011). IOP Publishing.
22. Zadeh, A. H., Alsabi, Q., Ramirez-Vick, J. E., & Nosoudi, N. (2020). Characterizing basal-like triple negative breast cancer using gene expression analysis: A data mining approach. *Expert Systems with Applications*, *148*, 113253.
23. Arunkumar, C., & 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), 131-142.
24. Alanni, R., Hou, J., Azzawi, H., & Xiang, Y. (2019). A novel gene selection algorithm for cancer classification using microarray datasets. *BMC medical genomics*, *12*(1), 1-12.
25. Mehmood, R., El-Ashram, S., Bie, R., Dawood, H., & Kos, A. (2017). Clustering by fast search and merge of local density peaks for gene expression microarray data. *Scientific reports*, *7*(1), 1-7.
26. Shukla, A. K. (2020). Multi-population adaptive genetic algorithm for selection of microarray biomarkers. *Neural Computing and Applications*, *32*(15), 11897-11918.