**EAI Endorsed Transactions**
**on Pervasive Health and Technology**

Research Article **EAI.EU**

# Efficient Gene Expression Data Analysis using ES-DBN For Microarray Cancer Data Classification

Swati Sucharita[1*], Barnali Sahu[2], Tripti Swarnkar[3]

[1,2] Department of Computer Science & Engineering, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, India
[3]Department of Computer Application, National Institute of Technology, Raipur, India

## Abstract

INTRODUCTION: DNA microarray has become a promising means for classification of various cancer types via the creation of various Gene Expression (GE) profiles, with the advancement of technologies. But, it is challenging to classify the GE profile since not all genes contribute to the presence of cancer and might lead to incorrect diagnoses. Thus an efficient GE data analysis for microarray cancer data classification using Exponential Sigmoid-Deep Belief Network (ES-DBN) is proposed in this work.
OBJECTIVES: The study aims to develop an efficient GE data analysis using Exponential Sigmoid-Deep Belief Network (ES-DBN) for microarray cancer data classification.
METHODS: The proposed methodology starts with pre-processing to compact data. Afterward, by utilizing Min-Max feature scaling technique, the pre-processed data is normalized. The normalized data is further encoded and feature ranking is performed. The subset values are selected using Cauchy Mutation-Coral Reefs Optimization (CM-CRO) in feature ranking. The feature vector is calculated by Pearson Correlation Coefficient based GloVe (PCC-GloVe) algorithm since different subsets return the same fitness value. Statistical and Biological validations take place after feature vector calculation. Lastly, for effective classification of the type of cancer, the vector features obtained are fed to ES-DBN.
RESULTS: The outcomes of the proposed technique are evaluated with various datasets, which exhibited that the proposed technique performed well with the Ovarian cancer dataset and outperforms other conventional approaches.
CONCLUSION: This study presents a comprehensive methodology for efficiently classifying cancer types using GE profile. The proposed GE data analysis using ES-DBN shows promising results, highlighting its potential as a valuable tool for cancer diagnosis and classification.

**Keywords:** Microarrays, Gene expression, cancer classification, Pearson Correlation Coefficient based GloVe (PCC- GloVe), Categorical columns, Exponential Sigmoid - Deep Belief Network (ES-DBN), Cauchy Mutation-Coral Reefs Optimization (CM-CRO).

Received on 11 01 2024, accepted on 15 05 2024, published on 28 05 2024

doi: 10.4108/ 10.4108/eetpht.10.6187

## 1. Introduction

Cancer, which causes huge death worldwide, is one of the deadliest diseases. The accurate prediction of the tumors type and size completely depends on the adoption of powerful and reliable classification models. Thus, the patients are provided with better treatment or else response to therapy [1]. Previously, cancer was classified and treated exclusively grounded on the organs of the origin or simplistic histomorphologic features [2]. Nowadays, the usage of GE phenotype in a variety of diagnostic areas for identification and classification purposes is improved owing to the advancement in technologies. The usage of microarrays leads to the monitoring of thousands of GEs per sample concurrently [3]. Microarray data-centric cancer classification assists in the effective early identification of cancer-causing genes, thus enhancing the lifetime of cancer patients [4].

*Corresponding author. Email: swatisucharita08@gmail.com

For the successful diagnosis and treatment of cancer, reliable as well as precise tumor classification is necessary. In the biomedical field, one of the most significant topics for cancer classification is gene selection [5]. Filter, wrapper, and embedded approaches are the three different approaches involved in the method of gene selection. The intrinsic characteristics of genes are collected in the filter approach that aids in the targeted phenotype class discrimination to directly select feature genes [6]. By utilizing the wrapper technique, the subset of feature genes is selected, which is then utilized for further assessment of new feature gene subsets. Thus, by utilizing this technique, the optimal number of feature genes can be selected automatically for a specific classifier. The embedded technique is the same as the wrapper technique in which to aid in the selection of feature subsets, multiple systems are combined in the embedded technique [7][8]. But, not all genes are cancerous genes and only a few genes are informative for each cancer type. The Hidden Markov Model (HMM) and dynamic programming model are wielded by most of the technologies that may incur an underflow problem [9]. When developing a cancer predictive model, one of the crucial issues is the GE profiles' higher dimensionality. For tumor classification, various techniques have been employed. Grounded on a variety of morphological, clinical, and molecular variables, the classification of human malignancies is involved in certain techniques. In other words, they cannot find an optimal solution in a reasonable time [10]. To overcome this issue, an efficient GE data analysis using ES-DBN for microarray cancer data classification is proposed; also, its main contributions are enlisted further,

- An efficient CM-CRO is introduced for feature subset selection.
- A novel PCC-GloVe is utilized for feature vector calculation.
- The ES-DBN in the classification step enhanced the classification accuracy and reduced the training time.

The remaining paper is systemized as: brief information on background work concerning tumor classification is elucidated in Section 2; the proposed technique is delineated in Section 3; the experimental outcomes are exemplified in Section 4; lastly, the paper is winded in Section 5.

## 2. Literature Survey

Lu et al., 2021 [11] propounded a hybridized Adaboost and Genetic Algorithm (GA) for the efficient classification of cancer utilizing GE data. Here, the formation of a decision group in the ensemble system enhanced the diversity of the classifiers. Hence, by avoiding the local extrema problem, the classification performance was enriched. On the contrary, the presented approach was affected by crossover and mutation rates.

Shukla et al., 2020 [12] established Teaching Learning-centric Optimization and Simulated Annealing (TLBOSA) for revealing the tumour's patterns as well as genes interpretability. Here, the optimal genes subsets were chosen; also, the redundant genes were filtered by Correlation-based Feature Selection (CFS). Afterward, to identify the most informative gene subsets, simulated annealing and TLBO algorithms were combined. Hence, the curse of dimensionality and the overfitting problem was avoided. However, the lack of exploitation was the major limitation.

Sampathkumar et al., 2020 [13] proffered the Cuckoo Search algorithm with Crossover (CSC) to select genes to aid in cancer classification. Here, the levy flight technique was wielded after the reproduction process. The outcomes validated the proposed mechanism's efficacy. But, the shortcoming here was the dependence of the optimization efficiency on the diversification and intensification strategy.

Algamal & Lee, 2019 [14] recommended a Two-Stage Sparse Logistic Regression (TS-SLR) to obtain an effective subset of genes with higher classification abilities. Here, the screening technique was wielded as a filter technique, whereas an adaptive lasso with a novel weight was utilized as an embedded mechanism in the initial stage. Thus, by utilizing this technique, the High correlation problem was avoided. Conversely, owing to design matrix singularity, logistic regression was infeasible.

L. Sun et al., 2019 [15] presented an entropy-centric gene selection technique with a Fisher score for tumor classification. Primarily, irrelevant genes were eliminated by the Fisher score technique. Then, for improving the classification performance, a joint neighborhood entropy-centric gene selection approach with the Fisher score was incorporated. Hence, better classification accuracy was obtained with lesser time complexity. Nevertheless, discretization led to the loss of certain beneficial information.

Sayed et al., 2019 [16] propounded an ensemble-based Nested GA (Nested-GA) for the selection of optimal feature subsets. Nested GA was made of inner and outer GA. Outer GA utilized a Support Vector Machine (SVM) for operating on GE data, whereas Neural Network (NN) structure was induced in inner GA for operating on DNA data. Therefore, by utilizing this technique, the problem of feature dependency was avoided. Yet, the classification accuracy was affected by the presence of outliers.

Mudiyanselage et al., 2020 [17] established Deep Fuzzy Neural Networks (DFNN) based framework for cancer classification. Firstly, the data was pre-processed and the informative genes were selected using a hybrid algorithm. Thereafter, by utilizing the deep fuzzy model, the noise and uncertainties were removed. Hence, the problem of data ambiguity was avoided. However, the presented approach was suitable only for the samples with smaller sizes.

## 3. Proposed Microarray Gene Data Classification

In solving GE profile problems, valuable outcomes are provided by microarray data analysis. But, since most of the genes are irrelevant or else insignificant to clinical diagnosis, classifying the GE profile is a challenging task. Thus, this work proposes an efficient GE data analysis using ES-DBN for microarray cancer data classification. Figure 1 elucidates the proposed technique's baseline structure.

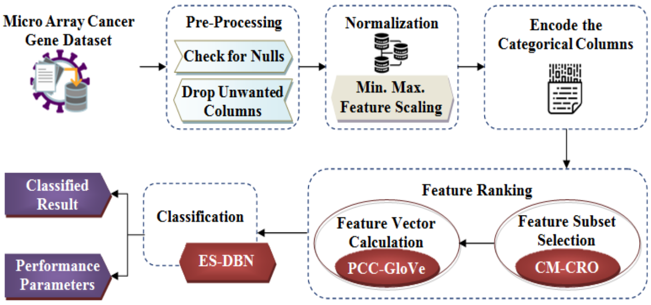


**Figure 1.** Block diagram of the proposed technique

### 3.1 Pre-processing

Primarily, the gene data obtained from the microarray dataset is pre-processed. In general, the GE data obtained from the microarray dataset contains more noise. Therefore, to minimize the uncertainties present in the data and provide accurate information for the better classification of the tumor types, pre-processing is mainly carried out. In the pre-processing stage, the two main steps involved are checking for nulls and dropping unwanted columns. It is detailed as follows.

- Check for nulls: Here, the gene data obtained from the dataset is checked for null values. Here, for effective classification results, the gene data with null values or the data with no attributes are removed. Therefore, the gene data obtained after checking for nulls is modeled as $G^n$.
- Drop unwanted columns: The dropping of unwanted columns takes place after null feature removal. As the gene data contains outliers, unwanted columns are removed (dropped) to improve classification accuracy by avoiding over-fitting problems. Hence, the resultant pre-processed gene data $\left(\overline{G}^n\right)$ are expressed as,

$$\overline{G}^n = \left\{\overline{G}^1, \overline{G}^2, ..., \overline{G}^g\right\} \tag{1}$$

Wherein, $n = 1,2,...g$ specifies the $g$-number of pre-processed data.

### 3.2 Normalization

Here, the pre-processed data is normalized to keep the pre-processed gene data values in a common range using the Min-Max feature scaling technique. Therefore, the expression for min-max normalization is expressed as,

$$\hat{D} = \frac{\overline{G}^n - \overline{G}^n_{\min}}{\overline{G}^n_{\max} - \overline{G}^n_{\min}} \tag{2}$$

Here, $\overline{G}^n_{\min}, \overline{G}^n_{\max}$ signify the maximum and minimum gene data, and $\hat{D}$ exemplifies the normalized data.

### 3.3 Encode the categorical columns

Here, the normalized data obtained from the normalization step comprises string (non-numerical) attributes, which will be then converted into numerical values for subsequent processing of gene data. Hence, the resultant encoded categorical columns are simplified as,

$$x_i = \{x_1, x_2, ...., x_n\} \tag{3}$$

Where, $i = 1,2,...,n$ symbolizes the number of encoded categorical columns.

### 3.4 Feature ranking

For effective tumor type classification, the feature ranking process is executed after encoding the categorical columns. Here, feature ranking is performed through feature subset selection and feature vector calculation phase and is further validated. The process is detailed further.

### 3.4.1 Feature Subset Selection via CM-CRO

In this step, by utilizing the CM-CRO algorithm, the selection of the most relevant subset features from the encoded categorical columns is carried out. CRO algorithm, which is mainly based on broadcast spawning, brooding, larvae setting, budding, and depredation operations, is a metaheuristic optimization algorithm. However, the randomized selection of coral reef couples in the broadcast spawning stage leads to local optimum problems with reduced convergence speed. To overcome this, the selection of the coral reef couples is done utilizing Cauchy Mutation (CM) technique. Hence, the usage of CM in the general CRO is named CM-CRO. CM-CRO steps are explicated further,

**Step 1:** Let $x_i = \{x_1, x_2, ...., x_n\}$ (encoded categorical columns) be the initial coral reef population, and $y_j = y_1, y_2, \cdots, y_p$ signifies the position of each coral reef in the $M \times N$ square grid.

**Step 2:** After population initialization, the performance of each coral reef is determined by calculating the fitness value (classification accuracy) that indicates the possible solution. Therefore, the fitness estimation process is mathematically represented as,

$$f(x_i) = f(x_1, x_2, ..., x_n) \tag{4}$$

Hence, the corals with a better fitness value $\left(f_x\right)$ survive longer than the others with poorer fitness values.

**Step 3:** After fitness evaluation, external sexual reproduction (Broadcast Spawning) takes place. Here, by utilizing the CM technique, certain coral reefs are selected as broadcast spawners. Thus, the broadcast spawners selected using the CM $\left(\varsigma^{x_i}\right)$is displayed as,

$$\varsigma^{x_i} = \frac{1}{\pi\alpha}\left[\frac{\alpha^2}{\left(f_x - \eta\right)^2 + \alpha^2}\right] \tag{5}$$

Wherein, $\alpha$ implies the parameter of scaling, and $\eta$ notates the location parameter.

**Step 4:** Next is the internal sexual reproduction stage (Brooding) stage. Here, hermaphrodite corals $\left(H^c\right)$ are produced through brooding. Therefore, the brooding stage is specified as,

$$H^c = 1 - f_x \tag{6}$$

Wherein, $\left(1 - f_x\right)$implies the fraction of coral reefs utilized in brooding.

**Step 5:** Next is the larvae setting phase. Here, the coral larvae $\left(H^c\right)$ produced via the internal or external sexual reproduction stages are set to the square grid and begin to grow. Before setting the coral larvae into the square grid, the fitness value of each coral larva is determined using (4).

**Step 6:** Corals can also be reproduced asexually through the budding or fragmentation process. Since the entire sexual coral reproduction process completely depends on a fitness function, certain corals with the fraction $\Re^{x_i}$ are duplicated and undergo the larvae setting the stage in asexual reproduction. Hence, the coral larvae reproduced asexually $\left(A^{\zeta}\right)$ are signified as,

$$A^{\zeta} = 1 - \Re^{x_i} \tag{7}$$

***Step*** *7*: Lastly, the end of the reproduction stage is depredation. Here, the larvae with poorer fitness values are depredated (killed) by other animals in the square grid (reef) to liberate space for the generation of corals. Therefore, the depredation process is modelled as,

$$D = \begin{cases} x_i > f\left(x_i\right), occupiesreef \\ x_i < f\left(x_i\right), depraded \end{cases} \tag{8}$$

Hence, the number of selected subset features $\left(S_f\right)$ is represented as,

$$S_f = \left(S_1, S_2, \cdots\cdots, S_m\right) \tag{9}$$

Here, the number of selected features is implied as $m$ . The pseudocode of the proposed CM-CRO is detailed further down.

**Pseudocode of CM-CRO**

```
Input: Encoded categorical columns $\{x_1, x_2, ...., x_n\}$
Output: Feature subset
Begin
        Initialize Coral Reef (CR), population size $(n)$ and maximum iteration $(I_{\max})$
    Set iteration $I = 1$
    While $(I \le I_{\max})$ do
        Calculate fitness
        Perform broadcast spawning with Cauchy mutation $(\varsigma^{x_i})$
        Perform Brooding with $1 - f_x$
        Perform larvae setting and $1 - \Re^{x_i}$
        Update local best CR $x_i$ by calculating fitness
        If $(x_i > f(x_i))$ {
                Occupy reef
        } Else if $(x_i < f(x_i))$ {
                Perform depradation
        }
        End If
     End while
     $I = I + 1$
     Return optimal feature subset $S_f$
End
```

## 3.4.2 Feature vector calculation by PCC-GloVe algorithm

Here, feature vectors are calculated using the Glove embedding algorithm for the obtained feature subsets. Glove embedding involves feature co-occurrence evaluation, defining soft constraints, and cost function stage. But, the co-occurrence evaluation grounded on the co-occurrence matrix solely relies on the matrix dimension, which in turn affects the classification accuracy. Therefore, Pearson Correlation Coefficient (PCC) is induced in the conventional GloVe approach to result in the PCC-GloVe algorithm and is detailed further.

- Initially, a vocabulary function $\left(V^k\right)$ is created to store the selected feature subsets and thereby compute the feature subset that occurs more frequently but without repeating. The mathematical formulation for vocabulary function is given as follows.

$$V^k = \sum_{f=1}^{m} \omega\left(S_f\right)\left(w_f^T w_f + u_f - \log S_f\right)^2 \tag{10}$$

Here, $\omega(S_f)$ specifies the weighting function, $w_f$ models the vector functions, and $u_f$ signifies the scalar biases. The weighting function is represented as,

$$\omega(S_f)=\begin{cases}\left(\dfrac{S_1}{S_m}\right)^{\delta}, \ if\ S_1 < S_m \\ 1\ otherwise\end{cases} \tag{11}$$

- Then, the co-occurrence of each feature subset $(\gamma)$ is determined by measuring the frequency of occurrence of each weighted feature subset using the PCC, and is formulated as,

$$\gamma=\frac{\sum\left(S_f-\bar{S}_f\right)\left(\omega(S_f)-\overline{\omega(S_f)}\right)}{\sqrt{\sum\left(S_f-\bar{S}_f\right)^2\sum\left(\omega(S_f)-\overline{\omega(S_f)}\right)^2}} \tag{12}$$

- Lastly, PCC-GloVe trains the dataset with the co-occurrences of feature subsets to provide a feature vector $(\upsilon^a)$ for each feature subset; hence, it will be stored inside the file for further processing. Therefore, the feature vectors $(\upsilon^a)$ calculated for each feature subset are referred as,

$$\upsilon^a=\upsilon^1,\upsilon^2,....,\upsilon^A \tag{13}$$

Where, $a=1,2,...,A$ implies the number of the feature vector. The feature vector thus obtained is validated for verifying the efficacy of the proposed approach.

### 3.4.3 Validation

For proving the proposed technique's worthiness, validation is performed. The vector features thus obtained are validated statistically and biologically and are detailed below.

- In statistical validation, the quality of the vector features obtained is evaluated through various statistical indices. Here, for the statistical analysis, ANOVA software is employed. ANOVA stands for Analysis of Variance, which provides a comparison of more samples. Here, the GE level of the microarray dataset is evaluated grounded on the weighted average of each gene among all samples.

- In biological validation, detailed information about the transfer functions is utilized in the proposed work, and the organ, which is affected by the tumor, is identified. In biological validation, a VOLCANO plot is utilized. It is a plot between log-transformed gene-specific t-tests and fold-change. Fold change refers to the ratio between the differentially expressed genes.

## 3.5 Classification by ES-DBN

For effective classification of the various types of tumors, the vector features thus obtained are fed into the ES-DBN classifier. A Deep Belief Network (DBN), which highly interacts with each other, is a neural network with a visible layer and a hidden layer. Since the activation function in DBN undergoes gradient diffusion problems, the Exponential Sigmoid (ES) activation function is incorporated to overcome this drawback in traditional DBNs. This modification in the conventional DBN's activation function is named ES-DBN. The stages involved in the ES-DBN are described further,

- Primarily, the feature vectors, $\upsilon^a=\upsilon^1,\upsilon^2,....,\upsilon^A$, are fed as input to the visible layer and are connected with the hidden layer provided with a hidden vector $\left(\hbar=\left\langle H_1,H_2,\cdots\cdots,H_h\right\rangle\right)$, and $h$ implies the number of the feature vectors in the hidden layer. The combined distribution of both visible and hidden layers is represented by an energy function $\left(\xi(\upsilon^a,\hbar)\right)$, which is formulated as,

$$\xi(\upsilon^a,\hbar)=-\rho^T\upsilon^a-\psi^T\hbar-\upsilon^{aT}\varpi\hbar \tag{14}$$

Wherein, the relationship betwixt the visible layer and hidden layer weights is notated as $\varpi$, the bias values of the visible and hidden layers are exemplified as $\rho,\psi$.

- Further, the Probability Density Function (PDF) of both visible as well as hidden layers $\left(\Gamma(\upsilon^a,\hbar)\right)$in terms of the energy function is given as,

$$\Gamma(\upsilon^a,\hbar)=\frac{1}{\sigma}e^{-\xi(\upsilon^a,\hbar)} \tag{15}$$

Here, $\sigma$ specifies the normalization constant and is detailed as,

$$\sigma=\sum_{\upsilon^a,\hbar}e^{-\xi(\upsilon^a,\hbar)} \tag{16}$$

- The conditional probability of the hidden layer $\left(\Gamma(\hbar=1\,|\,\upsilon^a)\right)$ is defined as,

$$\Gamma(\hbar=1\,|\,\upsilon^a)=\wp\left(\psi+\sum\varpi\hbar\right) \tag{17}$$

Likewise, the conditional probability of the visible layer $\left(\Gamma(\upsilon^a=1\,|\,\hbar)\right)$ is given as,

$$\Gamma(\upsilon^a=1\,|\,\hbar)=\wp\left(\rho+\sum\varpi\hbar\right) \tag{18}$$

In equation (17, 18), $\wp$ symbolizes the ES activation function and is expressed as,

$$\wp = \begin{cases} \upsilon^a , if\ \upsilon^a > 0 \\ \varepsilon\left(e^{\upsilon^a} - 1\right), \ otherwise \end{cases} \tag{19}$$

Here, $\varepsilon$ signifies the hyperparameter function.

- Thereafter, DBNs are fine-tuned by the backpropagation system. It means that the whole network's weight is adjusted continuously and is given in the following equation.

$$\Delta\varpi = \Im\left(\left\langle \upsilon^a , \hbar \right\rangle_{data} - \left\langle \upsilon^a , \hbar \right\rangle_{ran}\right) \tag{20}$$

Where, the learning rate is notated as $\Im$. Lastly, the tumor type is classified by the classifier, thus producing the outcomes in the output layer. Hence, the classifier's output is represented as,

$$r_d = \left[r_1, r_2, \cdots\cdots, r_D\right] \tag{21}$$

Where, $d = 1,2,..., D$ implies the different types of tumors. The pseudocode of the proposed ES-DBN is detailed below.

**Pseudocode of proposed ES-DBN**

```
Input: Features $\upsilon^a = \upsilon^1, \upsilon^2, ...., \upsilon^A$
Output: Classified output $r_d$
Begin
    Initialize visible layers, $\hbar$, $h$, bias $\rho, \psi$
    For feature $\upsilon^a$ do
        Perform combined distribution with $\left(\xi\left(\upsilon^a, \hbar\right)\right)$
        Estimate PDF with $\frac{1}{\sigma} e^{-\xi\left(\upsilon^a, \hbar\right)}$
        Estimate the conditional probability of visible layer by using ES activation
        Perform Fine tuning
        $\Delta\varpi = \Im\left(\left\langle \upsilon^a, \hbar \right\rangle_{data} - \left\langle \upsilon^a, \hbar \right\rangle_{ran}\right)$
        End For
        Return the output class $r_d$
End
```

## 4. Results and Discussion

Here, the detailed analysis of the proposed technique's outcome based on various performance metrics is explicated. The implementation of the proposed system is done by utilizing PYTHON along with an Intel i5/corei7 processor of 3.20GHz CPU speed using 4GB RAM and Windows 10 Operating System (OS). Here, the data are obtained from the publicly available dataset.

### 4.1 Dataset Description

Table 1 exemplifies the information regarding various datasets used in the proposed work. Here, 70% of the data is wielded for training, whereas 30% data for testing.

Table 1: Dataset Description

| Dataset | Sample | Genes | Class |
|---|---|---|---|
| Breast Cancer [18] | 151 | 54676 | 6 |
| Brain cancer [19] | 130 | 54676 | 5 |
| Acute lymphocytic leukemia (ALL), Acute myeloid leukemia (AML) [20] | 72 | 7130 | 2 |
| Colorectal [21] | 62 | 1935 | 4 |
| Leukemia [22] | 64 | 16384 | 5 |
| Lung cancer [23] | 181 | 1626 | 2 |
| Ovarian cancer [24] | 235 | 48 | 2 |

### 4.2 Performance Analysis of proposed ES-DBN without feature selection for different microarray datasets.

Here, the superiority of the proposed ES-DBN without feature selection is evaluated with numerous datasets and is displayed in Table 2.

Table 2: Performance analysis of proposed ES-DBN without feature selection

| Dataset | Accuracy | Precision | Sensitivity | Specificity | F-measure | MCC |
|---|---|---|---|---|---|---|
| Breast Cancer | 93.4782 | 92 | 95.8333 | 90.9090 | 93.8775 | 0.9077 |
| Brain Cancer | 92.3076 | 90.4761 | 95 | 89.4336 | 92.6829 | 0.9076 |
| ALL,AML | 86.3636 | 84.6153 | 91.6666 | 80 | 88 | 0.9072 |
| Colorectal cancer | 84.2105 | 81.8181 | 90 | 77.7777 | 85.7142 | 0.9085 |
| Leukemia | 85 | 83.3333 | 90.9090 | 77.7777 | 86.9565 | 0.9089 |
| Lung cancer | 94.5454 | 93.1034 | 96.4285 | 92.5925 | 94.7368 | 0.9097 |
| Ovarian cancer | 95.7746 | 94.5945 | 97.2222 | 94.2857 | 95.8904 | 0.9107 |

Table 2 exhibits that the ovarian cancer dataset attains better performance with the accuracy of 95.7746%, 94.5945% precision, 97.2222% sensitivity, 94.2857% specificity, 95.8904% f-measure, and 0.9107% MCC for classifying the tumor type. Also, the lung cancer dataset achieves accuracy, precision, sensitivity, specificity, f-measure, and MCC of the order of 94.5454%, 93.1034%, 96.4285%, 92.5925%, 94.7368%, and 0.9097%, correspondingly, which is lower when compared with ovarian cancer dataset. Likewise, the metric values obtained vary (lowers) for other datasets like breast cancer, brain cancer, cancer classification dataset, colorectal, and Leukemia dataset. Hence, the proposed ES-

DBN outstands the conventional approaches while utilizing the Ovarian cancer dataset.

## 4.3 Performance Analysis of proposed ES-DBN with feature selection for different microarray datasets.

Here, the performance of the proposed ES-DBN with feature selection is analyzed using numerous datasets and is explicated in Table 3.

Table 3: Performance evaluation of proposed ES-DBN with feature selection

| Dataset | Accuracy | Precision | Sensitivity | Specificity | F-measure | MCC |
|---|---|---|---|---|---|---|
| Breast Cancer | 95.6521 | 95.8333 | 95.8333 | 95.4545 | 95.8333 | 95.4545 |
| Brain Cancer | 94.8717 | 95 | 95 | 94.7368 | 95 | 94.7368 |
| ALL,AML | 90.9090 | 91.6666 | 91.6666 | 90 | 91.6666 | 90 |
| Colorectal | 89.4736 | 90 | 90 | 88.8888 | 90 | 88.8888 |
| Leukemia | 90 | 90.9090 | 90.9090 | 88.8888 | 90.9090 | 88.8888 |
| Lung cancer | 96.3636 | 96.4285 | 96.4285 | 96.2963 | 96.4285 | 96.2963 |
| Ovarian cancer | 97.1831 | 97.2222 | 97.2222 | 97.1428 | 97.2222 | 97.1428 |

Table 3 reveals the proposed technique's performance with feature selection based on the statistical measures using various datasets. The proposed ES-DBN system accomplishes higher accuracy of 97.1831% for the Ovarian cancer dataset, whereas the other datasets display accuracy values of 95.6521% (Breast cancer), 94.8717% (Brain cancer), 90.9090% (Cancer classification), 89.4736% (Colorectal), 90% (Leukemia), and 96.3636% (lung cancer), which are comparatively lower. Conversely, the precision, Sensitivity, Specificity, F-measure, and MCC of the Ovarian cancer dataset are 97.2222%, 97.2222%, 97.1428%, 97.2222%, and 97.1428%, correspondingly. However, the other datasets like Breast cancer, Brain cancer, Cancer classification, Colorectal, Leukemia, and lung cancer datasets show lower precision, sensitivity, specificity, F-measure, and MCC values. Therefore, it is clear that the utilization of the CM for selecting the features in the proposed ES-DBN technique provides better tumor classification when compared with Table 2 (without feature selection).

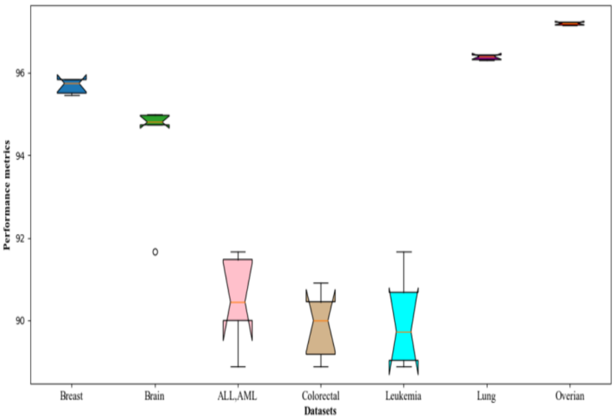
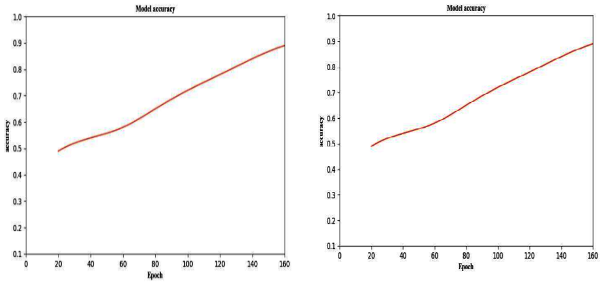


**Figure 2:** Graphical representation of proposed ES-DBN with feature selection

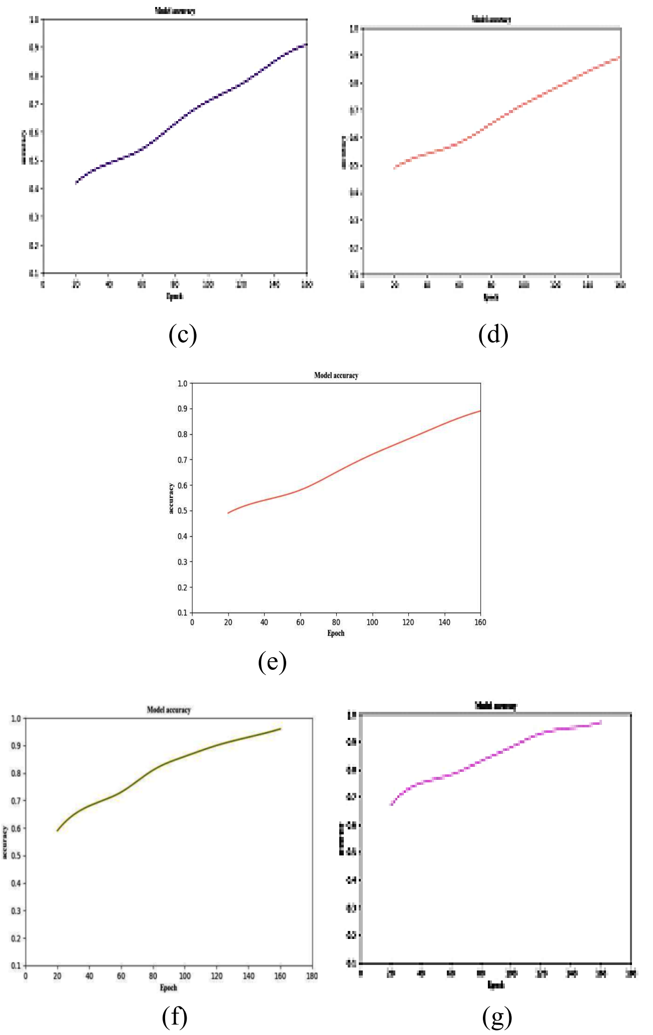
The graphical representation of the performance values achieved by the proposed ES-DBN using various datasets is elucidated in Figure 2. The accuracy, precision, Sensitivity, Specificity, F-measure, and MCC values obtained using the Ovarian dataset are 2.3%, 1.38%,1.38%,1.68%,1.38%, and 1.68% higher than the accuracy, precision, sensitivity, specificity, f-measure, and MCC of Breast cancer dataset. Likewise, the metric values obtained by the other dataset also differ (lower) when compared to the Ovarian dataset. Therefore, the proposed ES-DBN achieved better metrics rates. Thus, the proposed technique exactly classifies the type of tumor.

### 4.3.1 Performance evaluation of proposed ES-DBN based on convergence curve

Here, the proposed ES-DBN's superiority is evaluated in terms of the convergence curve using various datasets and is shown in Figure 3.



(a) (b)



(c) (d)

(e)

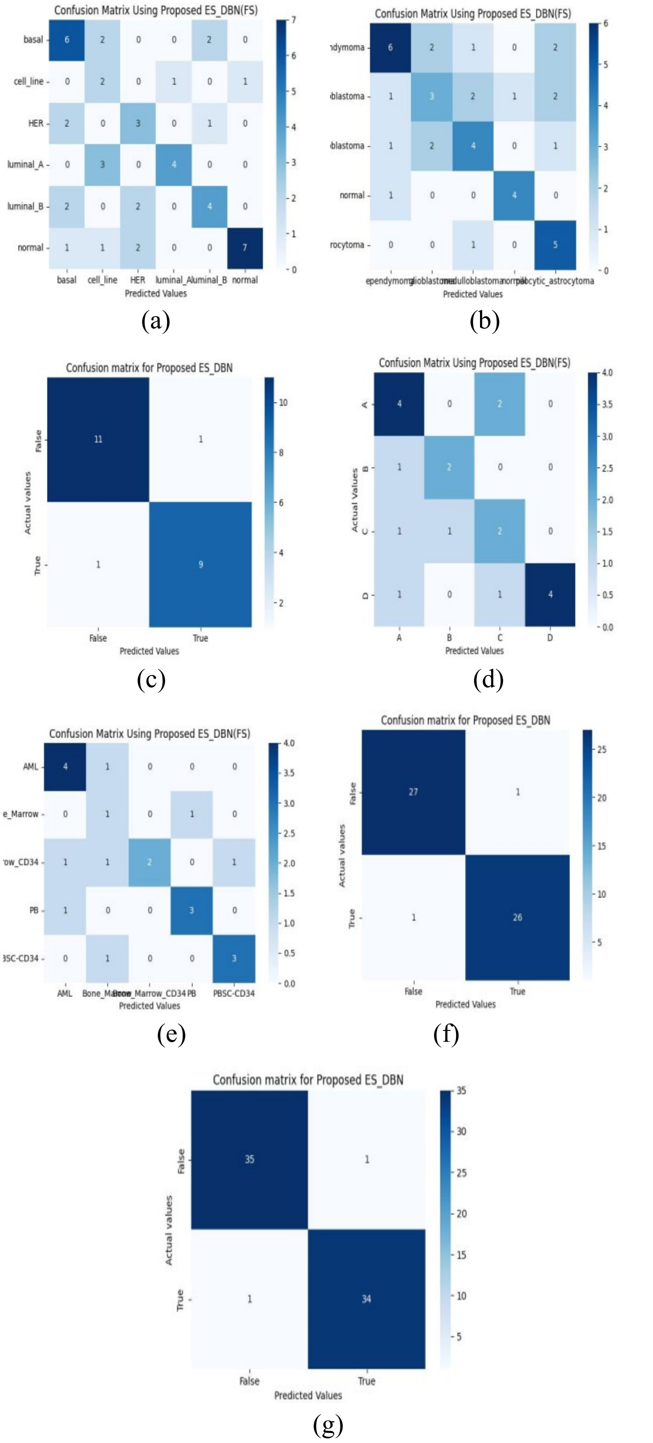
(f) (g)

**Figure 3:** Performance evaluation based on convergence curve for (a) Breast cancer gene expression (b) Brain cancer (c) ALL, AML dataset (d) Colorectal (e) Leukemia (f) Lung cancer (g) Ovarian cancer dataset

As per Figure 3, the proposed ES-DBN classifier begins the tumor-type classification process when the epoch value is 20 and completes the classification when the epoch value is 160. Among the various datasets used for analysis, the ovarian cancer dataset provided better classification accuracy of 97.1831% with an epoch value of 160, which is represented in figure 3 (g), whereas the breast cancer dataset attains 95.65217% accuracy (shown in figure 3 (a)) at the same 160th epoch value, which is lower than the Ovarian dataset. Likewise, the accuracy obtained by other datasets also varies and is exposed in Figure 3 (b, c, d, e, and f). Thus, the proposed work performs better with the Ovarian cancer dataset in the tumor type classification process.

### 4.3.2 Performance evaluation of proposed ES-DBN based on the confusion matrix

Figure 4 displays the prediction results regarding the tumor classification of the proposed approach in terms of the confusion matrix.



(a) (b)

(c) (d)

(e) (f)

(g)

**Figure 4:** Performance analysis of proposed ES-DBN in terms of a confusion matrix for (a) Breast cancer (b) Brain cancer (c) ALL, AML dataset (d) Colorectal (e) Leukemia (f) Lung cancer (g) Ovarian cancer dataset

The confusion matrix obtained by the proposed ES-DBN for different datasets is elucidated in Figure 4. Figure 4 exhibits that the ovarian cancer dataset truly predicted 35 normal classes and 34 abnormal classes and misclassified only 1 normal class as abnormal and 1 abnormal class as a normal class, whereas the breast cancer dataset truly predicted 6 basal as basal, and falsely predicted 2 basal as HER, 2 as Aluminal_B, and 1 as normal. Moreover, the number of HER truly predicted by the same breast cancer dataset is 3, whereas 2 are misclassified as Aluminal_B and 2 as normal, while the number of cell lines truly predicted is 2, and the misclassified value is 6, the number of Aluminal_A truly predicted is 4, and misclassified is 1, the number of Aluminal_B truly predicted is 4, and misclassified is 3, truly predicted as normal is 7 and falsely predicted is 1. Overall, the number of classes truly predicted by the breast cancer dataset is lower than its false prediction value. Similarly, the number of true and false predictions varies for other datasets but is not higher than for Ovarian datasets. Thus, the proposed methodology performs better using the ovarian dataset and aids in better classification.

## 4.4 Performance Analysis of proposed ES-DBN for different microarray datasets with various techniques.

Here, the proposed system's performance is validated with the existing Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), DBN, Artificial Neural Network (ANN), and Long Short-Term Memory (LSTM) techniques.

Table 4: Comparative evaluation of proposed ES-DBN with various techniques

| Dataset | Technique | Accuracy | Precision | Sensitivity | Specificity | F-measure | MCC |
|---|---|---|---|---|---|---|---|
| **Breast Cancer** | Proposed | 95.65217 | 95.8333 | 95.8333 | 95.4545 | 95.8333 | 95.4545 |
| | DBN | 93.4782 | 93.8775 | 93.8775 | 93.0232 | 95.8775 | 93.022 |
| | CNN | 86.9565 | 88.4615 | 88.4615 | 85 | 88.4615 | 85 |
| | ANN | 84.7826 | 88.4615 | 88.4615 | 80 | 86.7924 | 80 |
| | RNN | 89.1304 | 88.4615 | 88.4615 | 87.8048 | 90.1960 | 87.808 |
| | LSTM | 91.3043 | 92 | 92 | 90.4761 | 92 | 90.471 |
| **Brain cancer** | Proposed | 94.8717 | 95 | 95 | 94.7368 | 95 | 94.738 |
| | DBN | 92.3076 | 92.6829 | 92.6829 | 91.8918 | 92.6829 | 91.898 |
| | CNN | 84.6153 | 86.3636 | 86.3636 | 82.3529 | 86.3636 | 82.359 |
| | ANN | 82.0512 | 82.6087 | 86.3636 | 76.4705 | 84.4444 | 76.475 |
| | RNN | 87.1794 | 88.3720 | 88.3720 | 85.7142 | 88.3720 | 85.712 |
| | LSTM | 89.7435 | 90.4761 | 90.4761 | 88.8888 | 90.4761 | 88.888 |
| **ALL AML** | Proposed | 90.9090 | 91.6666 | 91.6666 | 90 | 91.6666 | 90 |
| | DBN | 86.3636 | 84.6153 | 91.6666 | 80 | 88 | 80 |
| | CNN | 72.7272 | 78.8714 | 78.5714 | 62.5 | 78.5714 | 62.5 |
| | ANN | 68.1818 | 73.3333 | 78.5714 | 50 | 75.8620 | 50 |
| | RNN | 72.2727 | 84.6153 | 78.5714 | 75 | 81.4814 | 75 |
| | LSTM | 81.8181 | 84.6153 | 84.6153 | 77.7777 | 84.6153 | 77.777 |
| **Colorectal** | Proposed | 89.4736 | 90 | 90 | 88.8888 | 90 | 88.888 |
| | DBN | 84.2105 | 85.7142 | 85.7142 | 82.3529 | 85.7142 | 82.359 |
| | CNN | 68.4210 | 75 | 75 | 57.1428 | 75 | 57.148 |
| | ANN | 63.1578 | 69.2307 | 75 | 42.8571 | 72 | 42.851 |
| | RNN | 73.6842 | 78.2608 | 78.2608 | 66.6666 | 78.2608 | 66.666 |
| | LSTM | 78.9473 | 81.8181 | 81.8181 | 75 | 81.8181 | 75 |
| **Leukemia** | Proposed | 90 | 90.9090 | 90.9090 | 88.8888 | 90.9090 | 88.888 |
| | DBN | 85 | 86.9565 | 86.9565 | 82.3529 | 86.9565 | 82.359 |
| | CNN | 70 | 76.9230 | 76.9230 | 57.1428 | 76.9230 | 57.148 |
| | ANN | 65 | 71.4285 | 76.9230 | 42.8571 | 74.0740 | 42.851 |
| | RNN | 75 | 80 | 80 | 66.6666 | 80 | 66.666 |
| | LSTM | 80 | 83.3333 | 83.3333 | 75 | 83.3333 | 75 |
| **Lung Cancer** | Proposed | 96.3636 | 96.4285 | 96.4285 | 96.2963 | 96.4285 | 96.293 |
| | DBN | 94.5454 | 93.1034 | 96.4285 | 92.5925 | 94.7368 | 92.595 |
| | CNN | 89.0909 | 90 | 90 | 88 | 90 | 88 |
| | ANN | 87.2727 | 87.0967 | 90 | 84 | 88.5245 | 84 |
| | RNN | 90.9090 | 93.1034 | 90 | 92 | 91.5254 | 92 |
| | LSTM | 92.7272 | 93.1034 | 93.1034 | 92.3076 | 93.1034 | 92.306 |
| **Ovarian cancer** | Proposed | 97.1831 | 97.2222 | 97.2222 | 97.1428 | 97.2222 | 97.148 |
| | DBN | 95.7746 | 94.5945 | 97.2222 | 94.2857 | 95.8904 | 94.287 |
| | CNN | 91.5493 | 92.1052 | 92.1052 | 909090 | 92.1052 | 90.900 |
| | ANN | 90.1408 | 89.7435 | 92.1052 | 87.8787 | 90.9090 | 87.877 |
| | RNN | 92.9577 | 94.5945 | 92.1052 | 93.9393 | 93.3333 | 93.933 |
| | LSTM | 94.3662 | 94.5945 | 94.5945 | 94.1176 | 94.5945 | 94.116 |

Concerning the performance metrics, the performance comparison of the proposed ES-DBN classifier with various prevailing techniques using different datasets is specified in Table 4. The proposed technique achieves better accuracy, precision, sensitivity, specificity, f-measure, and MCC for the Ovarian cancer dataset of the order of 97.1831%, 97.2222%, 97.2222%, 97.1428%, 97.2222%, and 97.1428%, correspondingly, while the existing methods like DBN conquer 95.7746% accuracy, 94.5945% precision, 97.2222% sensitivity, 94.2857% specificity, 95.8904% F-measure, 94.2857% MCC. Similarly, the performance metrics are lower for CNN, ANN, RNN, and LSTM. Therefore, it is clear that the usage of the ES function in DBN avoided the gradient divergence problem and aided in the better classification phenomenon when contrasted with conventional approaches.

## 4.5 Comparative measurement of proposed ES-DBN from literature papers

Here, concerning the accuracy, the proposed framework's performance is analyzed and contrasted with traditional methods like Adaboost-GA [11], TS-SLR [14], and Nested GA [16] and is revealed in Table 5.

Table: 5 Performance evaluation based on the accuracy

| Techniques | Accuracy(%) |
|---|---|
| Proposed ES-DBN | 95.53 |
| Adaboost-GA [11] | 95.33 |
| TS-SLR [14] | 96.35 |
| Nested GA [16] | 94.25 |

As per Table 5, the proposed ES-DBN increases the classification accuracy by attaining a higher range of accuracy (95.53%). In contrast to the proposed mechanism, the existing Adaboost-GA, TS-SLR, and Nested GA attained very lower rates of accuracy at the order of 95.33%, 96.35%, and 94.25%. Therefore, the proposed ES-DBN is superior to other conventional systems under various uncertain circumstances.

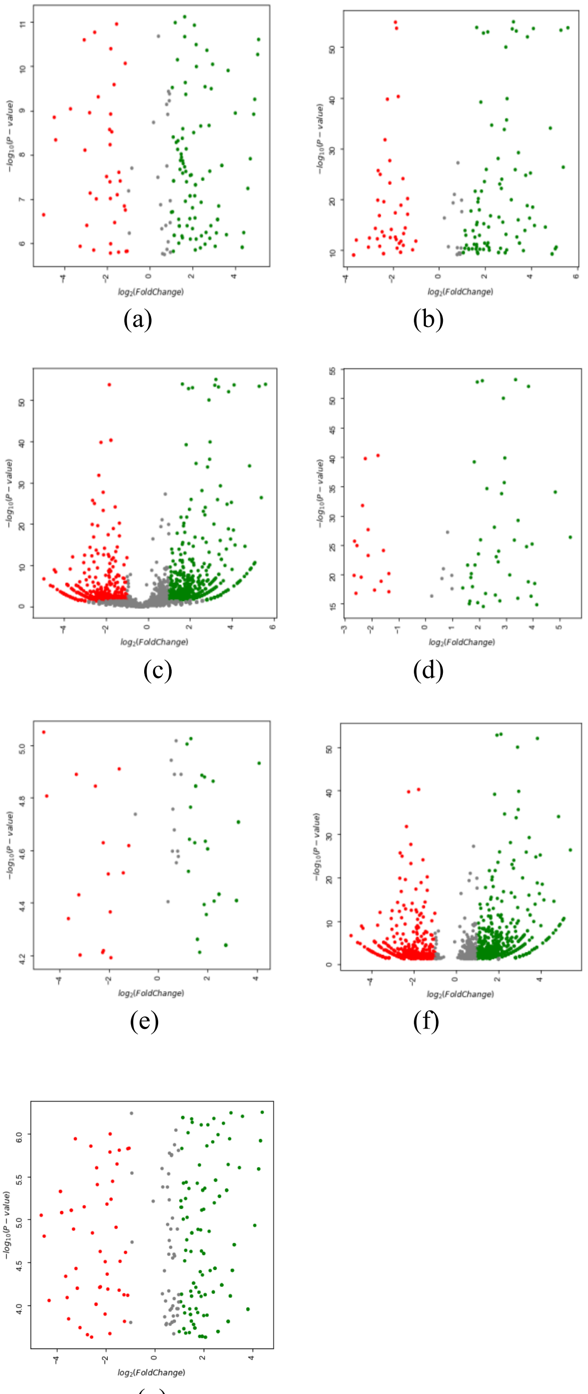
## 4.6 Biological Significant Analysis

The biological significant analysis is mainly carried out for analyzing the validated genes among the total number of genes selected for cancer classification.

Table 6: Biological significant analysis of proposed ES-DBN

| Dataset | Relevant number of features selected by model | No. of less significant genes | No. of more significant genes | No. of no significant genes |
|---|---|---|---|---|
| Breast Cancer | 35539 | 4975 | 27720 | 2844 |
| Brain cancer | 41202 | 4805 | 34047 | 2350 |
| ALL, AML | 4635 | 649 | 3615 | 371 |
| Colorectal | 1258 | 176 | 981 | 101 |
| Leukemia | 10635 | 4491 | 5738 | 406 |
| Lung cancer | 1057 | 148 | 824 | 85 |
| Ovarian cancer | 41 | 6 | 32 | 3 |

Table 6 exemplifies that among the 41 relevant features selected by ES-DBN using the Ovarian dataset, 6 genes are less significant, 32 genes are more significant, and only 3 genes are not significant for analysis. Likewise, the number of features selected and the number of more, less, and no significant genes obtained by the other datasets also vary but are not better than the Ovarian dataset. Therefore, the proposed technique performs better tumor classification using the ovarian dataset.



(a) (b)

(c) (d)

(e) (f)

(g)

**Figure 5:** Volcano plot-based biological significant analysis of proposed technique for (a)Breast cancer (b)Brain cancer (c) ALL, AML (d)Colorectal (e)Leukemia (f)Lung cancer (g) Ovarian cancer

Figure 5 elucidates the graphical representation of the biological significant analysis by means of a volcano plot. In the proposed work, a volcano plot is used for biological significant analysis as it identifies any changes in the GE data more fastly. Moreover, the genes with large-fold changes (most biologically significant genes) are also easily identified. Here, the most upregulated genes move toward the right, the less regulated genes move toward the left, and the most statistically significant genes move toward the top.

## 5. Conclusion

This paper proposed an efficient GE data analysis for microarray cancer data classification utilizing ES-DBN. The proposed technique undergoes various operations, namely Pre-processing, Normalization, Encoding of the categorical columns, Feature ranking, and Classification. Afterward, the experimental assessment is done, where the performance, as well as comparative evaluation, is executed concerning several metrics for validating the proposed mechanism's efficacy. For the analysis, various datasets are wielded, where the proposed technique performs better for the ovarian cancer dataset with and without feature selection by achieving 95.77465% of accuracy, 97.2222% of sensitivity, 94.2857% of specificity, 94.5945% precision, 95.89041% F-measure, and 0.9107% MCC without feature selection and 97.1831% of accuracy, 97.2222% of sensitivity, 97.1428% of specificity, 97.2222% precision, 97.2222% F-measure, and 97.1428% MCC using the feature selection step. Thus, the proposed efficient GE-based cancer data classification technique is superior to the conventional techniques; also, it remains to be more reliable and robust. The proposed work only evaluates the tumor type, but it fails to perform the risk assessment. Therefore, the work will be extended in the future with some advanced neural networks and optimization strategies to aid in the risk assessment process.

## REFERENCES

[1] Q. Liao, Y. Ding, Z. L. Jiang, X. Wang, C. Zhang, and Q. Zhang, "Multi-task deep convolutional neural network for cancer diagnosis," *Neurocomputing*, vol. 348, pp. 66–73, 2019, doi: 10.1016/j.neucom.2018.06.084.

[2] M. Ghosh, S. Begum, R. Sarkar, D. Chakraborty, and U. Maulik, "Recursive Memetic Algorithm for gene selection in microarray data," *Expert Syst. Appl.*, vol. 116, pp. 172–185, 2019, doi: 10.1016/j.eswa.2018.06.057.

[3] M. Mostavi, Y. C. Chiu, Y. Huang, and Y. Chen, "Convolutional neural network models for cancer type prediction based on gene expression," *BMC Med. Genomics*, vol. 13, no. Suppl 5, pp. 1–13, 2020, doi: 10.1186/s12920-020-0677-2.

[4] G. W. Wright *et al.*, "A Probabilistic Classification Tool for Genetic Subtypes of Diffuse Large B Cell Lymphoma with Therapeutic Implications," *Cancer Cell*, vol. 37, no. 4, pp. 551-568.e14, 2020, doi: 10.1016/j.ccell.2020.03.015.

[5] Y. Huo, L. Xin, C. Kang, M. Wang, Q. Ma, and B. Yu, "SGL-SVM: A novel method for tumor classification via support vector machine with sparse group Lasso," *J. Theor. Biol.*, vol. 486, p. 110098, 2020, doi: 10.1016/j.jtbi.2019.110098.

[6] A. Lopez-Rincon, M. Martinez-Archundia, G. U. Martinez-Ruiz, A. Schoenhuth, and A. Tonda, "Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection," *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–17, 2019, doi: 10.1186/s12859-019-3050-8.

[7] B. H. Shekar and G. Dagnew, "Grid search-based hyperparameter tuning and classification of microarray cancer data," *2019 2nd Int. Conf. Adv. Comput. Commun. Paradig. ICACCP 2019*, pp. 1–8, 2019, doi: 10.1109/ICACCP.2019.8882943.

[8] M. Daoud and M. Mayo, "A survey of neural network-based cancer prediction models from microarray data," *Artif. Intell. Med.*, vol. 97, pp. 204–214, 2019, doi: 10.1016/j.artmed.2019.01.006.

[9] S. P. Potharaju and M. Sreedevi, "Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance," *Clin. Epidemiol. Glob. Heal.*, vol. 7, no. 2, pp. 171–176, 2019, doi: 10.1016/j.cegh.2018.04.001.

[10] A. K. Shukla, D. Tripathi, B. R. Reddy, and D. Chandramohan, "A study on metaheuristics approaches for gene selection in microarray data: algorithms, applications and open challenges," *Evol. Intell.*, vol. 13, no. 3, pp. 309–329, 2020, doi: 10.1007/s12065-019-00306-6.

[11] H. Lu, H. Gao, M. Ye, and X. Wang, "A Hybrid Ensemble Algorithm Combining AdaBoost and Genetic Algorithm for Cancer Classification with Gene Expression Data," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 18, no. 3, pp. 863–870, 2021, doi: 10.1109/TCBB.2019.2952102.

[12] A. K. Shukla, P. Singh, and M. Vardhan, "A new hybrid wrapper TLBO and SA with SVM approach for gene expression data," *Inf. Sci. (Ny).*, vol. 503, pp. 238–254, 2019, doi: 10.1016/j.ins.2019.06.063.

[13] A. Sampathkumar, R. Rastogi, S. Arukonda, A. Shankar, S. Kautish, and M. Sivaram, "An efficient hybrid methodology for detection of cancer-causing gene using CSC for micro array data," *J. Ambient Intell. Humaniz. Comput.*, vol. 11, no. 11, pp. 4743–4751, 2020, doi: 10.1007/s12652-020-01731-7.

[14] Z. Y. Algamal and M. H. Lee, "A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification," *Adv. Data Anal. Classif.*, vol. 13, no. 3, pp. 753–771, 2019, doi: 10.1007/s11634-018-0334-1.

[15] L. Sun, X. Y. Zhang, Y. H. Qian, J. C. Xu, S. G. Zhang, and Y. Tian, "Joint neighborhood entropy-based gene selection method with fisher score for tumor classification," *Appl. Intell.*, vol. 49, no. 4, pp. 1245–1259, 2019, doi: 10.1007/s10489-018-1320-1.

[16] S. Sayed, M. Nassef, A. Badr, and I. Farag, "A Nested Genetic Algorithm for feature selection in high-dimensional cancer Microarray datasets," *Expert Syst. Appl.*, vol. 121, pp. 233–243, 2019, doi: 10.1016/j.eswa.2018.12.022.

[17] T. K. B. Mudiyanselage, X. Xiao, Y. Zhang, and Y.

Pan, “Deep Fuzzy Neural Networks for Biomarker Selection for Accurate Cancer Detection,” *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 12, pp. 3219–3228, 2020, doi: 10.1109/TFUZZ.2019.2958295.

[18] https://www.kaggle.com/datasets/brunogrisci/breast-cancer-gene-expression-cumida

[19] https://www.kaggle.com/datasets/brunogrisci/brain-cancer-gene-expression-cumida

[20] https://www.kaggle.com/datasets/crawford/gene-expression

[21] https://www.kaggle.com/code/docxian/colorectal-cancer-gene-expression-data-prep-eda/data -

[22] https://www.kaggle.com/datasets/brunogrisci/leukemia-gene-expression-cumida

[23] https://data.mendeley.com/datasets/ynp2tst2hh/4/files/f63db009-6ede-4484-9c11-804fb27af856

[24] https://www.kaggle.com/datasets/saurabhshahane/predict-ovarian-cancer