Computers, Materials & Continua
DOI: 10.32604/cmc.2022.027030
Article

Tech Science Press

Deep Learning Enabled Microarray Gene Expression Classification for Data Science Applications

Areej A. Malibari[1], Reem M. Alshehri[2], Fahd N. Al-Wesabi[3], Noha Negm[3], Mesfer Al Duhayyim[4], Anwer Mustafa Hilal[5,*], Ishfaq Yaseen[5] and Abdelwahed Motwakel[5]

[1]Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, 21589, Saudi Arabia
[2]Department of Information Technology, College of Computers and Information Technology, Taif University, Taif, 21944, Saudi Arabia
[3]Department of Computer Science, College of Science & Art at Mahayil, King Khalid University, Saudi Arabia
[4]Department of Natural and Applied Sciences, College of Community-Aflaj, Prince Sattam bin Abdulaziz University, Saudi Arabia
[5]Department of Computer and Self Development, Preparatory Year Deanship, Prince Sattam bin Abdulaziz University, AlKharj, Saudi Arabia
*Corresponding Author: Anwer Mustafa Hilal. Email: a.hilal@kku.edu.sa
Received: 09 January 2022; Accepted: 16 February 2022

Abstract: In bioinformatics applications, examination of microarray data has received significant interest to diagnose diseases. Microarray gene expression data can be defined by a massive searching space that poses a primary challenge in the appropriate selection of genes. Microarray data classification incorporates multiple disciplines such as bioinformatics, machine learning (ML), data science, and pattern classification. This paper designs an optimal deep neural network based microarray gene expression classification (ODNN-MGEC) model for bioinformatics applications. The proposed ODNN-MGEC technique performs data normalization process to normalize the data into a uniform scale. Besides, improved fruit fly optimization (IFFO) based feature selection technique is used to reduce the high dimensionality in the biomedical data. Moreover, deep neural network (DNN) model is applied for the classification of microarray gene expression data and the hyperparameter tuning of the DNN model is carried out using the Symbiotic Organisms Search (SOS) algorithm. The utilization of IFFO and SOS algorithms pave the way for accomplishing maximum gene expression classification outcomes. For examining the improved outcomes of the ODNN-MGEC technique, a wide ranging experimental analysis is made against benchmark datasets. The extensive comparison study with recent approaches demonstrates the enhanced outcomes of the ODNN-MGEC technique in terms of different measures.

Keywords: Bioinformatics; data science; microarray gene expression data classification; deep learning; metaheuristics

1 Introduction

Microarray classification and analysis are highly critical for earlier diagnoses and treatment of life-threatening diseases such as cancer. It displays the maximum rate of mortality and morbidity stands second in developing countries and in economically developed countries [1]. Generally, the human being suffers from two hundred kinds of cancer and the microarray technique is adapted for keeping records of them. The GLOBOCAN data, Global health observatory, United Nations World population prospectus, and World Health Organization reported that the four increasingly common cancer that occurs around the world are female breast, lung, prostate, and bowel cancer [2]. It is caused by oncogenes and is associated with genome. It causes uncontrolled and abnormal cell development.

The molecular examination makes known that distinct types of cancer have distinct gene expression profiles and might be used for diagnosing distinct cancers. Higher-density DNA microarray evaluates the activity of various genes in a similar manner. This novel technique assists in providing good therapeutic measurement to cancer patients by identifying type of cancer [3]. Earlier diagnosis of cancer types increases the possibility of survival for the victim. This diagnosis is frequently generated as a classification issue [4]. Therefore, it economically becomes restrictive to have larger sample sizes. In order to resolve these problems, microarray medicinal dataset needs dimensionality reduction [5].

There are two main problems facing the algorithm of microarray dataset: extreme amount of genes compared with a smaller amount of samples [6]. Although, there are wide-ranging techniques and algorithms are accessible for this higher dimension information, massive searching space (unrelated gene) damages the classification accuracy [7]. These unrelated genes confuse the learning method and are fed to unrelated genes that are prone to overfitting. A particular way to improve the accuracy of the classification with a higher dimension smaller sample dataset is gene selection (feature selection) [8]. Feature selection (FS) is the procedure of recognizing the related features from the data and representing the higher dimension dataset with a small searching space. However, for microarray data, the appropriate FS resolution is highly complicated as the sample size is smaller than number of genes. Various factors need to be considered when decreasing the dimensionality of the dataset [9]. The Two essential features are searching strategy and evaluation criteria of FS method. According to this factor, the FS method is separated as wrapper and filter methods [10].

In [11], a hybrid model based simulated annealing (SA) algorithm, adaptive neuro-fuzzy inference system (ANFIS), and fuzzy c-means clustering (FCM) are introduced. The presented approach is employed for classifying five distinct cancer data sets (that is central nervous system cancer, lung cancer, prostate cancer, brain cancer, and endometrial cancer). In [12], a Principal Component Analysis (PCA) reduction dimension approach involves the computation of proportion for eigenvector selection. For the classification model, a Levenberg-Marquardt Backpropagation (LMBP) and Support Vector Machine (SVM) approach have been chosen. The researchers in [13] presented a grid searching-based hyper parameter tuning (GSHPT) for RF parameter to categorize Microarray Cancer Data. A grid searching method is developed by a set of fixed parameters that is important in giving optimum performance based on n-fold cross-validation. The grid searching method offers optimal parameters includes several features to consider at all the splits, various trees in the forest. In this study, the ten-fold cross validation is taken into account.

The authors in [14] proposed a state-of-the-art Gene Selection Programming (GSP) approach to choose appropriate genes for efficient and effective classification of cancer. GSP depends on Gene Expression Programming (GEP) model with an improved mutation fitness function definition, recombination operators, and determined initial population. Support Vector Machine (SVM) with linear kernel serves as a classification of GSP. Sun et al. [15] presented an error correcting output code

(ECOC) method for classifying multiple class microarray dataset based data complexity (DC) model. In the study, an ECOC coding matrix is created according to the hierarchical partition of the class space by using Minimizing Data Complexity (ECOC-MDC).

This paper designs an optimal deep neural network based microarray gene expression classification (ODNN-MGEC) model for bioinformatics applications. The proposed ODNN-MGEC technique designs an Improved Fruit fly Optimization (IFFO) based feature selection technique that is utilized for reducing the high dimensionality in the biomedical data. Moreover, deep neural network (DNN) model is applied for the classification of microarray gene expression data and the hyperparameter tuning of the DNN model is carried out using the Symbiotic Organisms Search (SOS) technique. For examining the improved outcomes of the ODNN-MGEC technique, a wide ranging experimental analysis is made against benchmark datasets.

2 The Proposed Model

This paper has developed an ODNN-MGEC model for gene expression data classification in bioinformatics applications. The proposed ODNN-MGEC technique performs data normalization process to normalize the data into a uniform scale. Followed by, the IFFO algorithm is utilized for reducing the high dimensionality in the biomedical data. In addition, the DNN model is applied for the classification of microarray gene expression data and the hyperparameter tuning of the DNN model is carried out using the SOS algorithm. Fig. 1 demonstrates the overall block diagram of ODNN-MGEC technique.

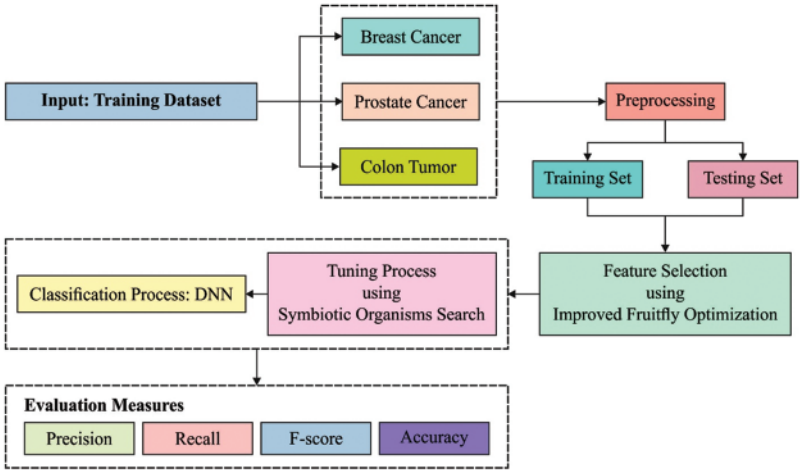

Figure 1: Block diagram of ODNN-MGEC technique

2.1 Data Normalization

In ML approaches were utilized for discovering tendencies from the dataset with comparative estimation amongst the dimensional data point. But endeavouring for using ML, an important issue was that there dimensional that are drastically varied scales. During this case, the min-max normalized was utilized for reducing the different scales of dimensional. The normalization alters the data from a particular small range by implementing linear transformation on original data. The dimensional value

of data is normalization from the range of zero and one utilizing min-max normalized. The min-max carries out the transformation of data by the subsequent formula as:

$$t = \frac{v - min_d}{max_d - min_d} (tran\_max_d - tran\_min_d) + tran\_min_d \quad (1)$$

where t refers the altered value of data value v from dimensional d, signifies the new minimal value and max_d implies the novel maximal value of dimensional d. Also, $tran\_min_d$ stands for the changed minimal value and $tran\_max_d$ signifies the transformed maximal value of dimensional d.

2.2 Algorithmic Design of IFFO Based Feature Selection

In this study, the proper election of feature subsets takes place via the IFFO algorithm. The FFO [16] is established dependent upon foraging performance of Drosophila. The fruit fly (FF) has higher to another species from olfactory ability and visual senses; so, it can be able to completely employ its drive for locating food. In detail, even at a distance of 40 km in the food sources, the nose of FF is collect different food scent which is dispersed during the air. With approaching the food source, the FFs place the food as well as companies gathering place with support of its sensitive visual organ, afterward, it is flying in that way. An optimum FF data are allocated with entire swarm under the iteration, and the next iteration is based only on data of preceding optimum FF. Based on the food search features of FF swarm, the FFO is separated as to many phases as follows [17]:

Step 1. Parameters initialized.

Initializing the parameter of FFO like maximal iteration number the population sizes, a primary FF swarm place ($X_$axis, $Y_$axis) and the arbitrary flight distance range.

$$X_{axis} = rands\,(1, 2) \quad (2)$$

$$Y_{axis} = rands\,(1, 2) \quad (3)$$

Step 2. Population initialized.

To provide the arbitrary place (X_i, Y_i) and distance to the food searching of individual FF, where i signifies the population sizes.

$$X_i = X\_axis + RandomValue \quad (4)$$

$$y_i = Y\_axis + RandomValue \quad (5)$$

Step 3. Population estimation.

Primarily, compute the distance of food place to origin (D). Next, calculate the smell focus judgment value (S) that is reciprocal of distance of the food place to origin.

$$D_i = \sqrt{X_i^2 + Y_i^2} \quad (6)$$

$$S_i = \frac{1}{D_i} \quad (7)$$

Step 4. Replacement.

Replacing the smell focus judgment value (S) with smell attention judgment function (is named as Fitness function) for finding the smell attention (Smell) of individual place of FF.

$$Smell_i = Function\,(S_i) \quad (8)$$

Step 5. Determine the higher smell attention.

Define the FF with maximum smell attention and the equivalent place amongst the FF swarm.

$$[bestSmellbestIndex] = \max(Smell) \tag{9}$$

Step 6. Retain the higher smell attention.

Recollect the maximum smell focus value and coordinates x and y. Afterward, the FF swarm fly near the place with high smell attention values.

$$Smellbest = bestSmell \tag{10}$$

$$X\_axis = X(bestIndex) \tag{11}$$

$$Y\_axis = y(bestIndex) \tag{12}$$

Step 7. Iterative optimization.

Enter the iterative optimized for repeating the execution of steps 2–5. The flow ends if the smell attention is not anymore higher than preceding iterative smell attention or if the iterative number attains the higher iterative numbers. The IFFO algorithm is derived by incorporating the concepts of chaos theory. The chaotic method is non-linear and divergent naturally, it illustrates optimum outcomes to global optimized. It creates oscillating trajectories and created a fractal infrastructure. The fitness function was resultant by IFFO technique to define solution in this state generated to obtain a balance among the 2 objectives as:

$$fitness = \alpha \Delta_R(D) + \beta \frac{|Y|}{|T|} \tag{13}$$

$\Delta_R(D)$ denotes the classifier error rate. $|Y|$ defines the size of subset that this method selects and $|T|$ whole count of features contained in the recent datasets. α demonstrates the parameter $\in [0, 1]$ relating to the weight of error rate of classification correspondingly however $\beta = 1 - \alpha$ implies the importance of reducing features.

2.3 Optimal DNN Based Classification

At the time of classification process, the DNN model is used to determine the proper class label. The basis of DNN is that the NN system is initially separated as a two-layer model and later train the two-layer NN system layer wise and lastly get the primary weight of multi-layer NN model by constructing the trained two-layer NN models, the entire procedure is named layer wise pre-training [18]. The hidden state of NN model extracts features from the input layer because of its abstraction. Therefore, the NN model using various hidden states is good at network generalization and processing as well attain fast convergence rate. Fig. 2 showcases the structure of DNN.

DNN is a kind of feed-forward ANN with many hidden states, also all the nodes at the similar hidden state utilize a similar non-linear function for mapping the input features from the layer below to the existing nodes. DNN framework is flexible because of the different hidden states and nodes, hence DNN illustrates outstanding capacity of fitting the complicated non-linear relations among inputs and outputs. In general, DNN method is employed for classification or regression. The relationships among inputs and outputs in DNN method is expressed by the following equation:

$$v^{l+1} = \rho\left(z^l\left(v^l\right)\right) \tag{14}$$

$$z^l\left(v^l\right) = w^l\left(v^l\right) + b^l, 0 \leq l < L \tag{15}$$

output $= v^L$

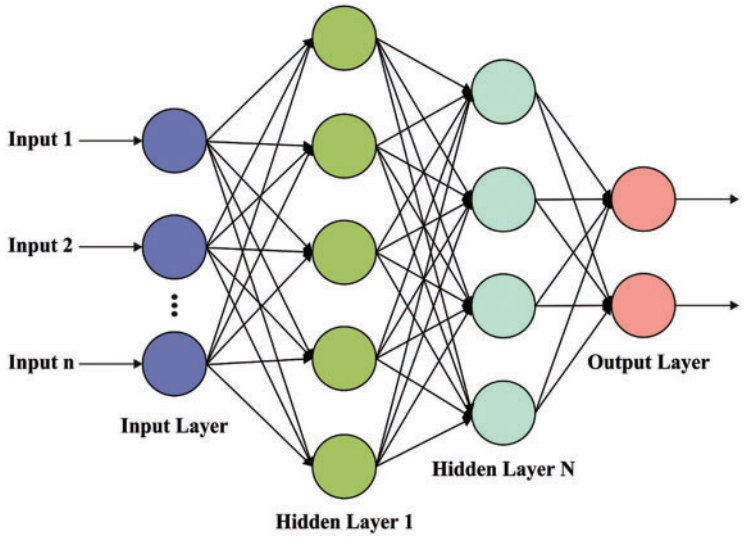

Figure 2: DNN structure

In the equation, we attain the last output by converting the feature vector of the initial layer v^0 to a processed feature vector v^l over L layers of nonlinear conversion. In the training phase, it is necessary to describe the offset vector b^l of l th layer and weight matrix w^l. With the variance among the actual and target outputs to generate a cost function, then, train DNN through backpropagation (BP) model. Mostly the proposal of DNN method comprises the number of nodes in all the layers, transfer function among the layers, the number of network layers, etc. The DNN method comprises the input, hidden, and output layers. In this work, the NN layer mostly focuses on identifying the amount of hidden states to describe the amount of layers. In NNs, hidden state has effects of abstraction and extracts feature from the input.

2.4 Hyperparameter Tuning Using SOS Algorithm

For optimally tuning the hyperparameters of the DNN model, the SOS is applied. Cheng and Prayogo [19] presented SOS, a novel population based metaheuristic approach simulated by natural ecosystems. SOS utilizes the symbiotic connection amongst 2 different species. The symbiotic connection that is general from the real world has mutualism, commensalism, and parasitism. Mutualism is described as inter-dependable connections amongst 2 organisms in which combined organism's advantage in the communication. The connection amongst the bee as well as flower is instance of mutualism connection. The bee moves amongst the flower and gather nectar and turned it as to honey. This activity profit the flower as it allows them from the pollination procedure. The procedure is expressed mathematically as:

$$P_i^{k+1} = P_i^k + rnd^{\star}\left(P_{best} - MV^{\star}BF1\right) \tag{16}$$

$$P_j^{k+1} = P_j^k + rnd^{*}\left(P_{best} - MV^{*}BF2\right) \tag{17}$$

where P_i implies the i^{th} member of populations and P_j refers the organisms that are chosen arbitrarily for interacting with P_i. Combined the organisms are functioning on mutual basis to survival from the ecosystems, *rnd* stands for the arbitrary number with uniform distribution amongst zero and one, MV signifies the mutual vectors, BF denotes the benefit vectors, k defines the generation and P_{best} represents the optimum individual organisms attained from k^{th} generation. MV and BF are computed as:

$$MV = \frac{P_i + P_j}{2} \tag{18}$$

$$BF = round\,(1 + rnd) \tag{19}$$

The round function has been utilized for setting the value of BF as one or two. BF has been utilized for identifying if the organism incompletely or completely benefits in the communication amongst individuals in the populations.

3 Experimental Validation

The proposed ODNN-MGEC technique has been validated using three datasets namely breast cancer, prostate cancer, and colon cancer datasets [20,21]. Breast cancer comprises 24,481 features and 97 samples. The prostate cancer dataset has 12,600 features with 136 samples where 77 samples are prostate tumors and 59 samples are normal. The colon cancer dataset has 2000 genes and 62 samples gathered in colon cancer patients.

Tab. 1 and Fig. 3 offer the experimental results obtained by the ODNN-MGEC technique on the breast cancer dataset. The results depicted that the ODNN-MGEC technique has offered enhanced classifier results under all hidden layers. For instance, with 2 hidden layers, the ODNN-MGEC technique has obtained $prec_n$, $reca_l$, $accu_y$, and $F1_{SCORE}$ of 66.81%, 66.82%, 81.88%, and 66.82% respectively. Likewise, with 10 hidden layers, the ODNN-MGEC system has reached $prec_n$, $reca_l$, $accu_y$, and $F1_{SCORE}$ of 63.28%, 69%, 82.76%, and 66.05% correspondingly. Similarly, with 20 hidden layers, the ODNN-MGEC methodology has achieved $prec_n$, $reca_l$, $accu_y$, and $F1_{SCORE}$ of 65.64%, 71.63%, 84.61%, and 67.61% correspondingly.

Table 1: Result analysis of ODNN-MGEC technique on breast cancer dataset

Breast cancer				
No. of hidden layers	Precision	Recall	Accuracy	F1-score
2	66.81	66.82	81.88	66.82
4	63.89	65.27	79.38	63.19
6	65.52	66.87	79.66	64.38
8	66.47	67.48	82.60	61.27
10	63.28	69.00	82.76	66.05
12	62.90	63.20	84.03	65.11
14	65.23	65.01	80.10	64.33
16	62.90	71.67	82.95	67.74
18	64.97	71.18	80.33	62.84
20	65.64	71.63	84.61	67.61

(Continued)

Table 1: Continued

Breast cancer				
No. of hidden layers	Precision	Recall	Accuracy	F1-score
Average	64.76	67.81	81.83	64.93

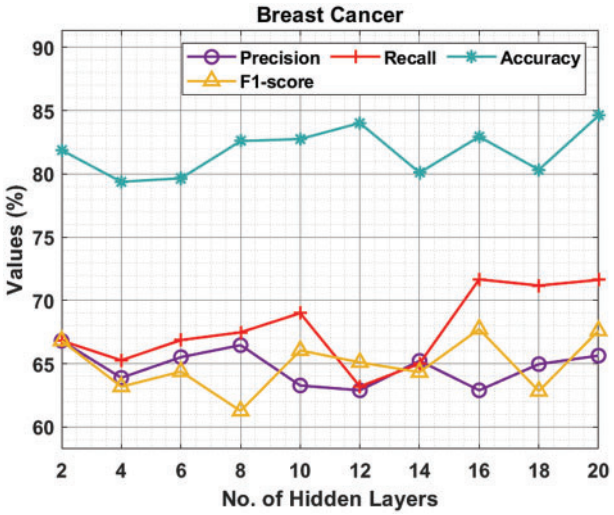

Figure 3: Result analysis of ODNN-MGEC technique on breast cancer dataset

Tab. 2 and Fig. 4 provide the experimental results obtained by the ODNN-MGEC technique on the prostate cancer dataset. The results depicted that the ODNN-MGEC technique has offered superior classifier outcomes under all hidden layers. For instance, with 2 hidden layers, the ODNN-MGEC method has obtained $prec_n$, $reca_l$, $accu_y$, and $F1_{SCORE}$ of 96.76%, 97.76%, 96.64%, and 96.63% respectively. Along with that, with 10 hidden layers, the ODNN-MGEC technique has obtained $prec_n$, $reca_l$, $accu_y$, and $F1_{SCORE}$ of 95.10%, 98.37%, 96.60%, and 97.62% correspondingly. Also, with 20 hidden layers, the ODNN-MGEC approach has obtained $prec_n$, $reca_l$, $accu_y$, and $F1_{SCORE}$ of 95.60%, 98.50%, 95.76%, and 96.90% correspondingly.

Table 2: Result analysis of ODNN-MGEC technique on prostate cancer dataset

Prostate cancer				
No. of hidden layers	Precision	Recall	Accuracy	F1-score
2	96.76	97.76	96.64	96.63
4	95.46	98.80	95.92	95.91
6	94.19	98.18	96.51	97.78
8	95.96	98.92	96.30	97.21
10	95.10	98.37	96.60	97.62
12	95.20	98.89	96.79	96.56

(Continued)

Table 2: Continued

Prostate cancer				
No. of hidden layers	Precision	Recall	Accuracy	F1-score
14	96.13	98.47	96.28	96.19
16	95.78	97.25	95.98	96.89
18	95.25	98.31	96.67	97.05
20	95.60	98.50	95.76	96.90
Average	95.54	98.35	96.35	96.87

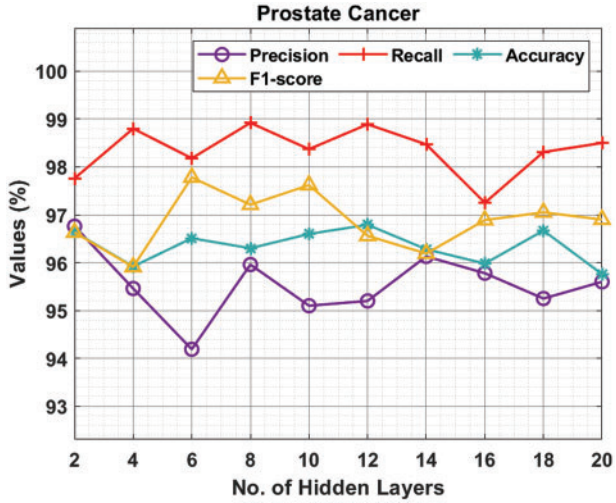

Figure 4: Result analysis of ODNN-MGEC technique on prostate cancer dataset

Tab. 3 and Fig. 5 demonstrate the experimental results obtained by the ODNN-MGEC system on the colon tumor dataset. The outcomes depicted that the ODNN-MGEC technique has obtainable improved classifier outcomes under all hidden layers. For instance, with 2 hidden layers, the ODNN-MGEC algorithm has achieved $prec_n$, $reca_l$, $accu_y$, and $F1_{SCORE}$ of 67.56%, 98.99%, 75.72%, and 82.84% correspondingly. Also, with 10 hidden layers, the ODNN-MGEC system has reached $prec_n$, $reca_l$, $accu_y$, and $F1_{SCORE}$ of 67.28%, 98.59%, 80.18%, and 81.94% correspondingly. In addition, with 20 hidden layers, the ODNN-MGEC technique has obtained $prec_n$, $reca_l$, $accu_y$, and $F1_{SCORE}$ of 71.20%, 98.61%, 81.40%, and 84.62% correspondingly.

Table 3: Result analysis of ODNN-MGEC technique on colon tumor dataset

Colon tumor				
No. of hidden layers	Precision	Recall	Accuracy	F1-score
2	67.56	98.99	75.72	82.84
4	70.80	98.74	80.93	79.99
6	70.82	98.19	75.07	83.08

(Continued)

Table 3: Continued

Colon tumor				
No. of hidden layers	Precision	Recall	Accuracy	F1-score
8	68.91	98.49	84.20	79.83
10	67.28	98.59	80.18	81.94
12	67.06	99.18	84.22	84.17
14	67.69	98.10	79.03	84.40
16	68.24	99.27	83.88	79.90
18	71.09	98.78	86.96	81.26
20	71.20	98.61	81.40	84.62
Average	69.07	98.69	81.16	82.20

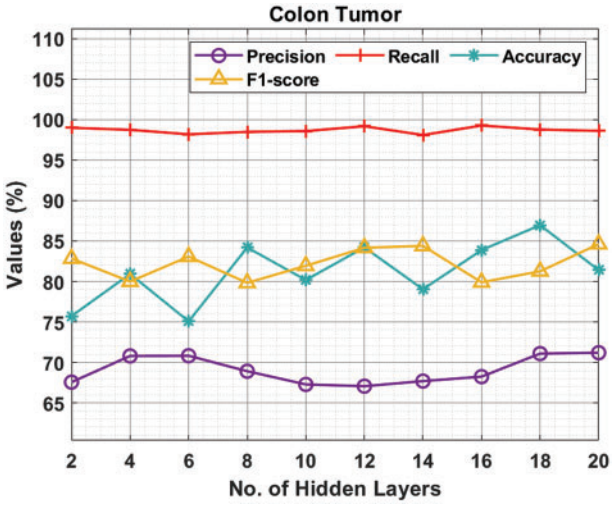

Figure 5: Result analysis of ODNN-MGEC technique on colon tumor dataset

In order to highlight the enhanced outcomes of the ODNN-MGEC technique, a comparison study is made with recent methods in Tab. 4.

Table 4: Comparative analysis of ODNN-MGEC technique with existing methods interms of different measures

Methods	Precision	Recall	Accuracy	F1- score
Breast cancer				
SVM-RBF	62.00	57.00	71.00	33.00
SVM-Polynomial	56.00	56.00	53.00	56.00
ODNN-MGEC	64.76	67.81	81.83	64.93
Prostate cancer				
SVM-RBF	93.00	98.00	95.00	96.00

(Continued)

Table 4: Continued

Methods	Precision	Recall	Accuracy	F1- score
SVM-Polynomial	90.00	32.00	50.00	47.00
ODNN-MGEC	95.54	98.35	96.35	96.87
Colon tumor				
SVM-RBF	68.00	98.40	70.00	81.00
SVM-Polynomial	62.00	76.00	55.00	68.00
ODNN-MGEC	69.07	98.69	81.16	82.20

Fig. 6 offers the classifier results of the ODNN-MGEC technique with existing techniques on breast cancer dataset. The results depicted that the SVM-Polynomial approach has resulted in poor performance with $prec_n$, $reca_l$, $accu_y$, and $F1_{SCORE}$ of 56%, 56%, 53%, and 56% respectively. Meanwhile, the SVM-RBF technique has attained slightly enhanced outcomes with $prec_n$, $reca_l$, $accu_y$, and $F1_{SCORE}$ of 62%, 57%, 71%, and 33% respectively. However, the ODNN-MGEC technique has outperformed the other methods with $prec_n$, $reca_l$, $accu_y$, and $F1_{SCORE}$ of 64.76%, 67.81%, 81.83%, and 64.93% respectively.

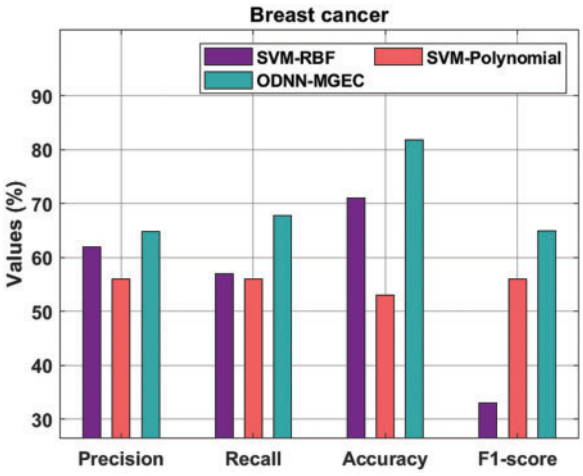

Figure 6: Comparative analysis of ODNN-MGEC technique on breast cancer dataset

Fig. 7 provides the classifiers of the ODNN-MGEC technique with existing techniques on prostate cancer dataset. The outcomes outperformed that the SVM-Polynomial approach has resulted in worse performance with the $prec_n$, $reca_l$, $accu_y$, and $F1_{SCORE}$ of 90%, 32%, 50%, and 47% correspondingly. In the meantime, the SVM-RBF method has obtained somewhat increased outcomes with $prec_n$, $reca_l$, $accu_y$, and $F1_{SCORE}$ of 93%, 98%, 95%, and 96% correspondingly. Eventually, the ODNN-MGEC approach has outperformed the other methods with $prec_n$, $reca_l$, $accu_y$, and $F1_{SCORE}$ of 95.54%, 98.35%, 96.35%, and 96.87% correspondingly.

Fig. 8 gives the classifier outcomes of the ODNN-MGEC methodology with existing algorithms on colon tumor dataset. The outcomes demonstrated that the SVM-Polynomial method has resulted in worse performance with $prec_n$, $reca_l$, $accu_y$, and $F1_{SCORE}$ of 62%, 76%, 55%, and 68% correspondingly. Followed by, the SVM-RBF methodology has gained somewhat superior outcomes with $prec_n$, $reca_l$,

$accu_y$, and $F1_{SCORE}$ of 68%, 98.40%, 70%, and 81% correspondingly. Lastly, the ODNN-MGEC technique has portrayed the other methodologies with $prec_n$, $reca_l$, $accu_y$, and $F1_{SCORE}$ of 62%, 76%, 55%, and 68% correspondingly.

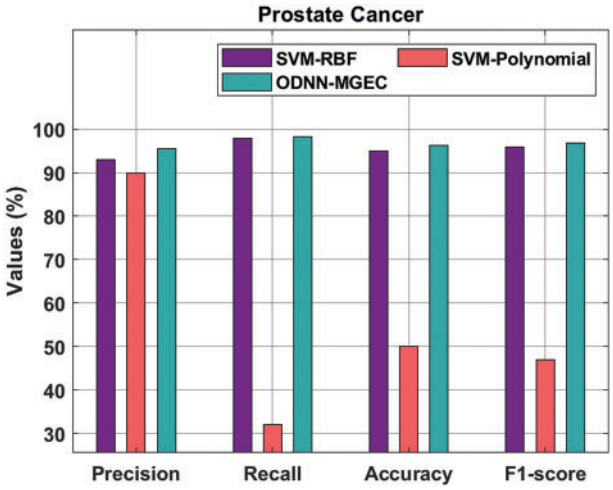

Figure 7: Comparative analysis of ODNN-MGEC technique on prostate cancer dataset

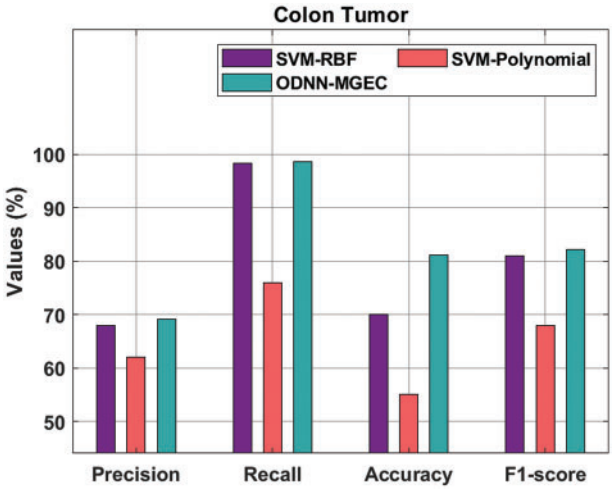

Figure 8: Comparative analysis of ODNN-MGEC technique on colon tumor dataset

From the above mentioned result and discussion, it is apparent that the ODNN-MGEC technique has accomplished superior outcomes over the other methods.

4 Conclusion

This paper has developed an ODNN-MGEC model for gene expression data classification in bioinformatics applications. The proposed ODNN-MGEC technique performs data normalization process to normalize the data into a uniform scale. Followed by, the IFFO algorithm is used to reduce the high dimensionality in the biomedical data. In addition, the DNN model is applied for the classification of microarray gene expression data and the hyperparameter tuning of the DNN

model is carried out using the SOS algorithm. The utilization of IFFO and SOS algorithms pave the way for accomplishing maximum gene expression classification outcomes. For examining the improved outcomes of the ODNN-MGEC technique, a wide ranging experimental analysis is made against benchmark datasets. The extensive comparison study with recent approaches demonstrates the enhanced outcomes of the ODNN-MGEC technique in terms of different measures. In future, the microarray gene classification performance can be boosted by the design of clustering and outlier removal models.

Funding Statement: The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work under grant number (RGP 2/42/43). This work was supported by Taif University Researchers Supporting Program (project number: TURSP-2020/200), Taif University, Saudi Arabia.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] R. Dash, "A two stage grading approach for feature selection and classification of microarray data using Pareto based feature ranking techniques: A case study," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 2, pp. 232–247, 2020.

[2] R. Dash, "An adaptive harmony search approach for gene selection and classification of high dimensional medical data," *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 2, pp. 195–207, 2021.

[3] P. Mohapatra, S. Chakravarty and P. K. Dash, "Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system," *Swarm and Evolutionary Computation*, vol. 28, no. Suppl. 8, pp. 144–160, 2016.

[4] S. Sucharita, B. Sahu and T. Swarnkar, "A comprehensivestudy on the application of grey wolf optimization for microarray data," in *Data Analytics in Bioinformatics: A Machine Learning Perspective*, pp. 211–248, 2021.

[5] R. Dash and B. B. Misra, "Pipelining the ranking techniques for microarray data classification: A case study," *Applied Soft Computing*, vol. 48, no. 11, pp. 298–316, 2016.

[6] C. Zhang, L. Liu, S. Zhang and C. Huang, "Microarray data classification based on neighbourhood components analysis projection method," in *2021 IEEE 6th Int. Conf. on Big Data Analytics (ICBDA)*, Xiamen, China, pp. 123–127, 2021.

[7] S. Begum, R. Sarkar, D. Chakraborty, S. Sen and U. Maulik, "Application of active learning in DNA microarray data for cancerous gene identification," *Expert Systems with Applications*, vol. 177, no. 11, pp. 114914, 2021.

[8] R. Dash and B. Misra, "Gene selection and classification of microarray data: A Pareto DE approach," *Intelligent Decision Technologies*, vol. 11, no. 1, pp. 93–107, 2017.

[9] Jahwar and N. Ahmed, "Swarm intelligence algorithms in gene selection profile based on classification of microarray data: A review," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 01–9, 2021.

[10] R. Dash and B. B. Misra, "Performance analysis of clustering techniques over microarray data: A case study," *Physica A: Statistical Mechanics and its Applications*, vol. 493, no. 1, pp. 162–176, 2018.

[11] B. Haznedar, M. T. Arslan and A. Kalinli, "Optimizing ANFIS using simulated annealing algorithm for classification of microarray gene expression cancer data," *Medical & Biological Engineering & Computing*, vol. 59, no. 3, pp. 497–509, 2021.

[12] U. N. Wisesty Adiwijaya, E. Lisnawati, A. Aditsania and D. S. Kusumo, "Dimensionality reduction using principal component analysis for cancer detection based on microarray data classification," *Journal of Computer Science*, vol. 14, no. 11, pp. 1521–1530, 2018.

[13] B. H. Shekar and G. Dagnew, "Grid search-based hyperparameter tuning and classification of microarray cancer data," in *2019 Second Int. Conf. on Advanced Computational and Communication Paradigms (ICACCP)*, Gangtok, India, pp. 1–8, 2019.
[14] R. Alanni, J. Hou, H. Azzawi and Y. Xiang, "A novel gene selection algorithm for cancer classification using microarray datasets," *BMC Medical Genomics*, vol. 12, no. 1, pp. 10, 2019.
[15] M. Sun, K. Liu, Q. Wu, Q. Hong, B. Wang *et al.,* "A novel ECOC algorithm for multiclass microarray data classification based on data complexity analysis," *Pattern Recognition*, vol. 90, pp. 346–362, 2019.
[16] W. T. Pan, "A new fruit fly optimization algorithm: Taking the financial distress model as an example," *Knowledge-Based Systems*, vol. 26, no. 7, pp. 69–74, 2012.
[17] L. Shen, H. Chen, Z. Yu, W. Kang, B. Zhang *et al.,* "Evolving support vector machines using fruit fly optimization for medical data classification," *Knowledge-Based Systems*, vol. 96, no. 3, pp. 61–75, 2016.
[18] W. Zheng, D. Hu and J. Wang, "Fault localization analysis based on deep neural network," *Mathematical Problems in Engineering*, vol. 2016, pp. 1–11, 2016.
[19] M. Y. Cheng and D. Prayogo, "Symbiotic organisms search: A new metaheuristic optimization algorithm," *Computers & Structures*, vol. 139, pp. 98–112, 2014.
[20] A. H. Chen and C. Yang, "The improvement of breast cancer prognosis accuracy from integrated gene expression and clinical data," *Expert Systems with Applications: An International Journal*, vol. 39, no. 5, pp. 4785–4795, 2012.
[21] B. Liu, Q. Cui, T. Jiang and S. Ma, "A combinational feature selection and ensemble neural network method for classification of gene expression data," *BMC Bioinformatics*, vol. 5, pp. 136, 2004.