IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 3, 2020, accepted January 20, 2020, date of publication January 29, 2020, date of current version February 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2970210

Artificial Intelligence Technique for Gene Expression by Tumor RNA-Seq Data: A Novel Optimized Deep Learning Approach

NOUR ELDEEN M. KHALIFA[1], MOHAMED HAMED N. TAHA[1], DALIA EZZAT ALI[1], ADAM SLOWIK[2], (Senior Member, IEEE), AND ABOUL ELLA HASSANIEN[1]

[1]Faculty of Computers and Artificial Intelligence, Cairo University, Giza 12613, Egypt
[2]Department of Electronics and Computer Science, Koszalin University of Technology, 75-453 Koszalin, Poland

Corresponding author: Adam Slowik (aslowik@ie.tu.koszalin.pl)

ABSTRACT Cancer is one of the most feared and aggressive diseases in the world and is responsible for more than 9 million deaths universally. Staging cancer early increases the chances of recovery. One staging technique is RNA sequence analysis. Recent advances in the efficiency and accuracy of artificial intelligence techniques and optimization algorithms have facilitated the analysis of human genomics. This paper introduces a novel optimized deep learning approach based on binary particle swarm optimization with decision tree (BPSO-DT) and convolutional neural network (CNN) to classify different types of cancer based on tumor RNA sequence (RNA-Seq) gene expression data. The cancer types that will be investigated in this research are kidney renal clear cell carcinoma (KIRC), breast invasive carcinoma (BRCA), lung squamous cell carcinoma (LUSC), lung adenocarcinoma (LUAD) and uterine corpus endometrial carcinoma (UCEC). The proposed approach consists of three phases. The first phase is preprocessing, which at first optimize the high-dimensional RNA-seq to select only optimal features using BPSO-DT and then, converts the optimized RNA-Seq to 2D images. The second phase is augmentation, which increases the original dataset of 2086 samples to be 5 times larger. The selection of the augmentations techniques was based achieving the least impact on manipulating the features of the images. This phase helps to overcome the overfitting problem and trains the model to achieve better accuracy. The third phase is deep CNN architecture. In this phase, an architecture of two main convolutional layers for featured extraction and two fully connected layers is introduced to classify the 5 different types of cancer according to the availability of images on the dataset. The results and the performance metrics such as recall, precision and F1 score show that the proposed approach achieved an overall testing accuracy of 96.90%. The comparative results are introduced, and the proposed method outperforms those in related works in terms of testing accuracy for 5 classes of cancer. Moreover, the proposed approach is less complex and consume less memory.

INDEX TERMS Cancer, RNA sequence, deep convolutional neural network, gene expression data.

I. INTRODUCTION

Cancer Cancer is a general term that used to describe a group of diseases associated with abnormal cell growth with metastatic and invasive characteristics [1]. In 2018, cancer was responsible for more than 9 million deaths worldwide. Approximately 17% of females and 20% of males will have cancer at some point in time, and 10% of females and 13% of males will die from it [2]. Based on statistics from the WHO, every year, more than 8 million people die from cancer, accounting for approximately 13% of deaths worldwide, indicating that cancer is one of the most threatening diseases in the world [1]. In 2018, lung cancer (1.76 million deaths) and colorectal cancer (860,000) are recorded as the most common cancers. Stomach cancer (780,000), liver cancer (780,000), and breast cancer (620,000) ranked second, third and fourth among the most common cancers [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Victor Hugo Albuquerque.

A tumor is any irregular cell proliferation that may be either benign or malignant. A benign tumor remains limited to its original location, which does not invade normal tissue or spread to distant locations of the body. Nevertheless, a malignant tumor may invade normal tissue and spread throughout the body through the circulatory or lymphatic systems (metastasis). The majority of cancers are classified into one of three major groups: carcinomas, sarcomas, and leukemias or lymphomas. Carcinoma is a type of cancer that develops from epithelial cells, accounting for 90% of cancers in human. Histological forms of carcinoma include adenocarcinoma, squamous cell carcinoma, adenosquamous carcinoma, anaplastic carcinoma and large cell carcinoma.Grading of carcinomas refers to the employment of criteria intended to semi-quantify the degree of cellular and tissue maturity seen in the transformed cells relative to the appearance of the normal parent epithelial tissue from which the carcinoma derives. The grades vary from grade 1 to grade 4 [1].

The Cancer Genome Atlas (TCGA) is a landmark genomics cancer data set including gene expression, DNA methylation, somatic mutation, copy number variation, microRNA expression, and protein expression. These expressions are sequenced and molecularly characterized over 11,000 cases of primary cancer samples. A joint project was established between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) under the name of TCGA. In 2006, TCGA was treated as a pilot project. Its focus was on three cancer types: lung, ovarian, and glioblastoma. In 2009, NHGRI and NCI reauthorized TCGA for a complete production phase, due to the significant success of the initial efforts. In the following decade, TCGA collected more than 11,000 cases across 33 tumor types and generated a large, comprehensive database describing the molecular changes that occur in tumours [3], [4]. Those large datasets provided a great classification opportunity for the global landscape of aberrations at RNA, DNA and protein levels [5].

This paper proposes an optimized deep learning approach based on BPSO-DT and CNN to classify normal and tumor conditions depending on a high-dimensional RNA-Seq gene expression data. For a high level of accuracy in classification, the high-dimensional RNA-seq data has been optimized with BPSO-DT to reduce its dimensions by selecting only the best features and removing the irrelevant features. Then to input the optimized RNA-seq results into the CNN architecture, the results were embedded into 2-D images. To avoid overfitting, different data augmentation techniques have been applied to the 2D images. The proposed approach was trained and evaluated on a public RNA-seq dataset consists of five separate cancer types, namely kidney renal clear cell carcinoma (KIRC), uterine corpus endometrial carcinoma (UCEC), breast invasive carcinoma (BRCA), lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC).

The remainder of this paper is organized as follows. Section II presents the methods and materials that discuss the background of this work, while Section III explores related work. Section IV discusses the dataset. Section V illustrates the proposed approach, while Section VI discusses our outcomes and issues in the paper. Finally, Section VII provides conclusions and directions for further research.

II. METHODS AND MATERIALS

A. BINARY PARTICLE SWARM OPTIMIZATION

Particle swarm optimization (PSO) is a stochastic optimization technique developed by Kennedy and Eberhart [6]. PSO is a population-based search algorithm based on the organism's behavior on a social milieu, of which a bird flock or a fish school are representative examples [7]. PSO involves simple mathematics and does not require high computational speeds. Moreover, it uses a small number of parameters to adjust and similar parameters can be used for various continuous optimization problems and also the discrete optimization problems such as the feature selection problem [7]. In PSO, every single solution of the target problem is represented by a particle. A group of particles is called a swarm. The whole swarm flies in the D-dimensional search space to find the optimal solutions by updating the position of each particle based on the experience of its own and its neighboring particles. All particles have fitness values, which evaluated using the fitness function, and have velocities that guide the movement of the particles. During movement, the current position of particle i at k iteration is denoted by a vector $X_i^k = (x_{i1}, x_{i2}, \ldots, x_{iD})$ and the velocity of particle i at k iteration is denoted as $V_i^k = (v_{i1}, v_{i2}, \ldots, v_{iD})$. Each particle updates its velocity and position depending on two fitness value are the local fitness value (Pbest) and the global fitness value (Gbest) according to equations (1), (2).

$$V_i^k = wV_i^{k-1} + c_1 r_1 \left(Pbest_i - X_i^{k-1}\right) + c_2 r_2 \left(Gbest - X_i^{k-1}\right) \tag{1}$$

$$X_i^k = X_i^{k-1} + V_i^k \tag{2}$$

where w is inertia weight, c_1 and c_2 are acceleration constants, r_1 and r_2 are random numbers uniformly distributed between 0 and 1.

PSO was originally introduced to solve continues problems. However, there are very discrete problems such as the feature selection problem. Therefore, Kennedy and Eberhart extended PSO to binary PSO (BPSO) [8] to solve discrete problems. In BPSO, equation (1) is still used to update the velocity, where X_i, $Pbest_i$ and $Gbest$ are limited to take the values 1 or 0, for this reason, the position update equation becomes a probabilistic equation. A sigmoid function $sig(V_i^k)$ is used to transform the V_i^k to the range of (0,1) as shown in equation (3). According to the update mechanism in BPSO, each particle updates each position based on the particle's velocity, which acts as probability threshold as shown in the probabilistic equation (4).

$$sig\left(V_i^k\right) = \frac{1}{1 + e^{-_i^k}} \tag{3}$$

$$X_i^k = \begin{cases} 1, & \text{if } rand < sig(V_i^k) \\ 0, & \text{if } otherwise \end{cases} \tag{4}$$

where 1 means this feature is chosen as an important feature for the next regeneration and 0 means this feature is not chosen as an important feature to the next regeneration, and rand is a random number $\in [0, 1]$.

B. DEEP LEARNING

Traditional image processing techniques provided reasonable results and performance in medical disease detection using infected and uninfected images, but it was limited to small data sets and theoretical results. As deep learning has revolutionized the area of computer vision [9]–[11], specifically image detection and object classification and recognition, it is now considered as a promising tool to improve such automated diagnosis systems to achieve higher outcomes, widen disease scope, and perform applicable real-time medical imaging [12]–[17] for disease classification systems.

Deep learning (DL), as a branch of Artificial Intelligence, depends on algorithms for data processing and thinking process simulation or for developing abstractions [18]–[20]. DL use layers of algorithms to process, analyse and discover hidden patterns in data and human speech understanding, and visually objects recognition [21]–[23]. Information passed through each layer of a deep network, with the output of the previous layer providing input for the next layer. The input layer is the first layer in the network, while output layer is the last layer in the network. All the layers located between the input and output layers are referred to as hidden layers of the network. Each layer is typically a simple, uniform algorithm containing one kind of activation function [18], [24].

C. CONVOLUTION NEURAL NETWORKS

Until 2011, CNN analysis was not prominent at computer vision conferences and journals, but in June 2012, a paper by Ciregan *et al.* [25] at the leading conference CVPR showed how max-pooling CNNs on GPU can dramatically improve many vision benchmark records. In October 2012, a similar system introduced by Krizhevsky *et al.* [26] won the large-scale ImageNet [27] competition by achieving a significant classification accuracy margin over classical ma-chine learning methods. In November 2012, Ciresan et al.'s [28] system also won the ICPR contest on the analysis of large medical images for cancer detection and in the following year, it won the MICCAI Grand Challenge on the same topic.

In the following years, various advances in deep CNNs further reduced the error rate on the ImageNet task. Several representative CNNs like VGGNet [29], GoogLeNet [30], and Residual Neural Network (ResNet) [31] demonstrated significant improvements in successive ImageNet Large-Scale Visual Recognition Competition (ILSVRC) annual challenges. A model called Xception [32] was introduced that uses depth-wise separable convolutions to outperform the Inception-V3 model [33] on the ImageNet [27] dataset classification task. A new CNN variant called densely connected convolutional networks (DenseNet), introduced by Huang *et al.* [34], utilizes a network architecture in which each layer is directly connected to every later layer. DenseNet has achieved noticeable improvements over the state-of-the-art while using significantly fewer parameters and computations.

III. RELATED WORK

This section conducts a survey on the latest studies for applying deep learning and machine learning in the field of tumor gene expression data. Researchers worldwide have begun to apply machine and deep learning tools to obtain significant results in a wide variety of medical image analyses/understanding tasks. In [35], Hsu et al. uses RNA sequencing data from The Cancer Genome Atlas (TCGA), and they focus on classifying 33 types of cancer patients. The authors introduced five machine learning algorithms, namely, DT, KNN, linear support vector machine (linear SVM), polynomial support vector machine (poly SVM), and an artificial neural network (ANN). The best result shows that linear SVM is the best classifier in this study, with a 95.8 Lyu and Haque [36] designed a new method to discover potential biomarkers for each tumor type. Based on the pan-cancer atlas, the method was provided with abundant information on 33 prevalent cancer tumor types. They used a convolutional neural network to classify tumor types and used a visualization neural network method to discover top tumor genes from the input. The high-dimensional RNA-Seq data was embedded into 2-D images and was used as a convolutional neural network to make classification of the 33 cancer tumor types. Based on the idea of Guided Grad Cam, as to each class, they generated a significance heat-map for all the genes. The proposed system achieved 95.59% using a train/test split.

The authors [5] undertook the development of a pan-cancer atlas to recognize 9,096 TCGA tumor samples representing 31 tumor types. They randomly assigned 75% (approximately 6800 samples) of samples into the training set and 25% (approximately 2300 samples) into the testing set, proportionally allocating samples from each tumor type. For the non-sex-specific tumor classification, they eliminated all tumor types that are sex-specific, namely, BRCA, CESC, OV, PRAD, TGCT, UCEC, and UCS. For the remaining tumor types, the samples were separated into two groups based on the patient gender. Three additional tumor types (CHOL, DLBC, and KICH) were eliminated due to small gender-specific sample sizes. The authors applied the genetic algorithm and k-nearest neighbours (KNN) methods to iteratively generate the subset of the genes (features) and then use the KNN method to test the accuracy. This method achieved an accuracy of 90% across 31 tumor types and generated a set of top genes for all the tumor types.

TABLE 1. Units for Magnetic Properties.

Tumor Type	BRCA	KIRC	LUAD	LUSC	UCEC
Number of Samples	878	537	162	240	269

The deep learning method was also used to classify top tumor genes and identify individual cancer types. In paper [37], the authors first used a stacked denoising autoencoder (SDAE) to extract high-level features from high-dimensional gene expression profiles. The authors then input these features into a single-layer ANN network to decide whether the sample is a cancerous tumor or not. The accuracy using this method reached 94%. The results and analysis illustrate that these highly interactive cancer genes could be useful for the detection of breast cancer.

Xiao *et al.* [38] presented a semi-supervised deep learning strategy called the stacked sparse auto-encoder (SSAE) to classify and predict cancer tumor using RNA-seq data. The proposed SSAE-based method employs the pre-training gredey layer approach and a sparsity penalty term to capture and extract important information from the high-dimensional data and then classify the samples. The proposed SSAE model was tested on three public RNA-seq data sets of three types of cancers, lung adenocarcinoma (LUAD), stomach adenocarcinoma (STAD) and breast invasive carcinoma (BRCA). They compared the prediction performance with several commonly used classification methods. The proposed SSAE-based semi-supervised learning model achieves the best classifications of 98.15%, 96.23%, and 99.89% for the STAD, BRCA and LUAD datasets, respectively. Xiao *et al.* [39] also demonstrated a new strategy, which used deep learning for an ensemble approach that incorporates multiple different machine learning models. The proposed deep learning-based multi-model ensemble method was applied to three public RNA-seq datasets representing three kinds of cancers, lung adenocarcinoma (LUAD), stomach adenocarcinoma (STAD) and breast invasive carcinoma (BRCA). It obtains improved predictions of 99.20%, 98.41%, and 98.78% for the LUAD, BRCA, and STAD datasets, respectively.

IV. TUMOR GENE EXPRESSION DATASET

The tumor gene expression dataset used in this research was published in [40]. It consisted of the RNA sequencing values from tumor samples belonging to five separate cancer types: lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), breast invasive carcinoma (BRCA), kidney renal clear cell carcinoma (KIRC) and uterine corpus endometrial carcinoma (UCEC). This dataset contains 2,086 rows and 972 columns, each row contains a specific sample, and each column contains the RPKM RNA-Seq values of a specific gene. The last column contains the cancer categories encoded numerically: 1 = BRCA, 2 = KIRC, 3 = LUAD, 4 = LUSC, 5 = UCEC. The number of samples for each tumor category is illustrated in Table 1.

V. PROPOSED ARCHITECTURE

The proposed architecture consists of three phases. The first phase is the pre-processing, while the second phase is the data augmentation. Deep learning training is the third phase, which relies on the deep N.

Algorithm 1 Extraction the Important Features of RNA-Seq

```
Input : Tumor gene expression dataset
Output: Gbest position
Initialize the position and velocity of each particle randomly
while iteration condition is not satisfied do
    Evaluate the fitness of the particle swarm by DT according to equation 5
    for each particle i do
        if the fitness of $x_i$ is greater than the fitness of the $Pbest_i$ then
            $Pbest_i = x_i$
        end
        if the fitness of any particle of the swarm is greater than Gbest then
            Gbest = particle's position
        end
        for each dimension $D = 1, \ldots, N$ do
            Update particles velocity and particles position according to equation 1,3, and 4 respectively
        end
    end
    go to next generation until termination criterion is met
end
Output Gbest
```

A. PRE-PROCESSING PHASE (BPSO-DT AND 2D IMAGE CREATION)

In this phase, BPSO is applied to implement the feature selection, and the decision tree (DT) [8] is used as BPSO's fitness function for a classification problem. In the context of this work, BPSO is used to reduce the number of RNA-seq features to a minimum and select only important features, and increase the accuracy of the classification. Therefore, the fitness function is calculated as equation (5) [41].

$$Fitness = \alpha\left(1 - C_p\right) + (1 - \alpha)\left(1 - \frac{S_f}{T_f}\right) \tag{5}$$

where α is a hyperparameter that decides the tradeoff between the classifier performance C_p, and the size of the feature subset S_f with respect to the total number of features T_f. In this work the classifier performance is the accuracy.

The steps of BPSO-DT are presented in Algorithm 1, where the input is the tumor gene expression dataset consisting of RNA-seq that needs optimization. By following all steps 1 through 16 mentioned in Algorithm 1, the Gbest position representing the best specific features of RAN-seq is returned. During this experiment, the parameters of BPSO were set as follows: the maximum number of iterations = 10, number of particles = 600, $c_1 = c_2 = 1.5$, $w = 0.7$, $v_{max} = 6$, $v_{min} = -6$, and $D = 971$ which is the dimension of the dataset used.

In BPSO-DT processing, 615 features out of 971 features were chosen as the best features of RNA-seq. Then, a set of steps have been applied to the optimized tumor gene expression dataset to convert it from data format to image format. The preprocessing phase include 1) loading the tumor gene expression on memory, 2) Change the data numerical domain range from [0, 24248] to image range [0, 255] according to equation (6), 3) Construct image by converting the optimized data record of 615 cells into a 25 × 25 pixels image. The result of this phase will be 2086 images classified into 5 tumor categories. Figure 1 illustrates a set of images after the preprocessing phase.

$$PixelValue = Round\left(\frac{Cell\ Value\ *\ 255}{24248}\right) \tag{6}$$

where 24248 is the maximum cell value in the tumor gene expression data, and 255 is the maximum value of the image domain.

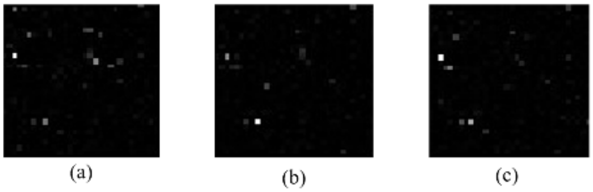


FIGURE 1. Magnetization as a function of applied field. The output of the pre-processing phase for KIRC (a), BRCA (b), and LUSC (c).

B. DATA AUGMENTATION PHASE

The proposed architecture for the deep learning model, which will be presented in the section on the following phase, has a huge number of learnable parameters compared to the number of images in the training set. The original dataset after the preprocessing phase contains 2086 images for 5 classes of tumor gene expression. Because of the great difference between learnable parameters and the number of images in the training set, the model is very likely to overfit. Deep learning models perform better with large datasets. One very widespread way to make datasets bigger is data augmentation or jittering. Data augmentation can increase the size of the dataset up to 10 or 20 times the original one or more, which helps avoid overfitting when training on very little data. The approach assists in building simpler and robust models which can generalize better. In this section, the common techniques that have been used in this research for overcoming the overfitting problem are presented.

C. AUGMENTATION TECHNIQUES

The most common method to overcome overfitting is to increase the number of images used for training by applying label-preserving transformations. In addition, data augmentation schemes are applied to the training set to make the resulting model more invariant to reflection, zooming and small noise in pixel values. To apply augmentation, each image in the training data is transformed as follows:

- *Reflection around X axis,*
- *Reflection around Y axis,*
- *Reflection around X-Y axis,*
- *Zooming.*

The mentioned data augmentation techniques have been applied to the dataset, this raises the dataset a number of images from 2086 images to 10430 images, doubled 5 times. This will lead to a significant improvement in the neural network training phase. Additionally, will make the proposed deep learning architecture immune to memorize the data and be more robust and accountable for the testing phase.

D. DEEP LEARNING TRAINING PHASE

This research conducted many experimental trails to propose the following deep learning architecture. Those experiments were already implemented in previous studies in [42]–[46] but the testing accuracy was unacceptable. Therefore, there was a need to propose a new one.

The proposed deep learning architecture of tumor gene expression is introduced in detail in Figure 2 and Figure 3. A graphical representation of the proposed architecture is shown in Figure 2. Figure 3 illustrates the layer details of the proposed architecture. The architecture consists of 14 layers, including two convolutional layers for features extraction with different convolution window 3*3 pixels, followed by two fully connected layers for classification.

The first layer is the input layer with input size 25*25 pixels. The second layer is the convolution layer with window size 3*3 pixels and 16 different filters. The third layer is a ReLU, which is used as the nonlinear activation function, then followed by an intermediate pooling with subsampling in layer four. A convolution layer with window size 3*3 pixels and 32 different filters and ReLU activation function are applied in the sixth and seventh layers. A dropout layer is in layer number eight to overcome the overfitting problem. Then, layer nine is a fully connected layer with 64 neurons, with a ReLU activation function. The last fully connected layer has 5 neurons to classify 5 classes for the tumor gene expression in layer number thirteen and uses a softmax layer to obtain the class memberships.

VI. EXPERIMENTAL RESULTS

The proposed architecture was developed using a software package (MATLAB). The implementation was GPU specific. All experiments were performed on a computer server with an Intel Xeon E5-2620 processor (2 GHz), 32 GB of RAM and 12 GB Nvidia GTX Titan X.

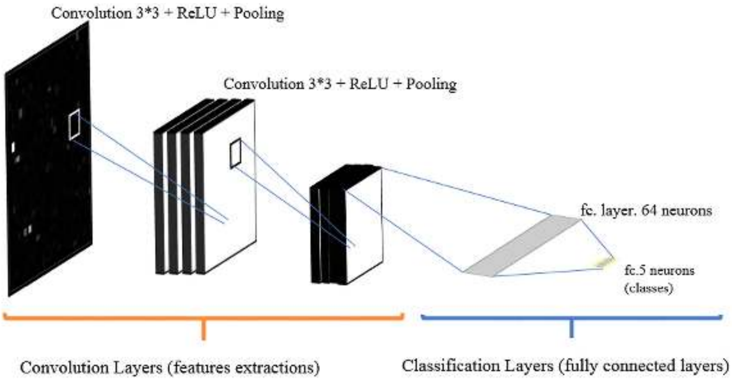

Convolution 3*3 + ReLU + Pooling
Convolution 3*3 + ReLU + Pooling
fc. layer. 64 neurons
fc.5 neurons
(classes)
Convolution Layers (features extractions)
Classification Layers (fully connected layers)

FIGURE 2. Magnetization as a function of applied field. Graphical representation of the proposed deep learning CNN architecture for tumor gene expression classification.

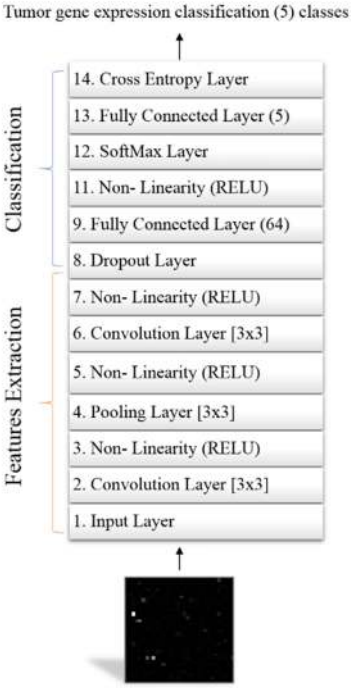
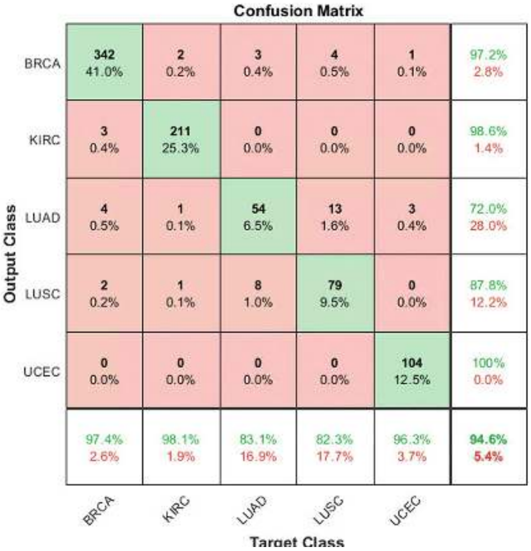


FIGURE 3. Detailed layer description for the proposed deep learning CNN architecture for tumor gene expression classification.

A. TESTING ACCURACY MEASUREMENT

To measure the accuracy of the proposed architecture for tumor gene expression using deep convolutional neural networks, 5 different trials were performed, and the median accuracy was calculated for different training and testing splitting. This research adopted three splitting strategies to evaluate the proposed architecture. The first strategy is splitting the data into 60% for training and 40% for testing, while the second strategy is splitting the data into 70% for training and 30% for testing. The last strategy is splitting the data into 80% for training and 20% for testing. The confusion matrices for the different strategies are presented in Figure 4, 5, and 6.

Figure 4, 5, and 6 illustrates that the more data is used for training phase the more accuracy the model will achieve.



	BRCA	KIRC	LUAD	LUSC	UCEC	
BRCA	342 41.0%	2 0.2%	3 0.4%	4 0.5%	1 0.1%	97.2% 2.8%
KIRC	3 0.4%	211 25.3%	0 0.0%	0 0.0%	0 0.0%	98.6% 1.4%
LUAD	4 0.5%	1 0.1%	54 6.5%	13 1.6%	3 0.4%	72.0% 28.0%
LUSC	2 0.2%	1 0.1%	8 1.0%	79 9.5%	0 0.0%	87.8% 12.2%
UCEC	0 0.0%	0 0.0%	0 0.0%	0 0.0%	104 12.5%	100% 0.0%
	97.4% 2.6%	98.1% 1.9%	83.1% 16.9%	82.3% 17.7%	96.3% 3.7%	94.6% 5.4%

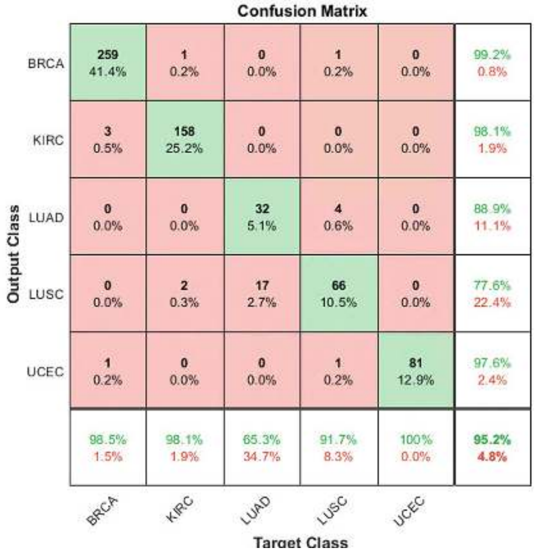
FIGURE 4. Confusion matrix for 60% training and 40% testing strategy.

TABLE 2. Median testing accuracy for different training and testing strategies using BPSO-DT or without using BPSO-DT.

Training and Testing Strategy	Strategy 2 (60% -40%)	Strategy 3 (70% -30%)	Strategy 4 (80% -20%)
Median Testing Accuracy without BPSO-DT	93.20%	94.40%	96.20%
Median Testing Accuracy without BPSO-DT	94.60%	95.20%	96.90%

Table 2 summarizes the achieved median accuracy for the different adopted training and testing strategies with/without BPSO-DT. The proposed deep learning architecture for tumor gene expression had achieved median testing accuracy with 96.90% in the 80% training and 20% testing strategy with BPSO-DT, which means that the using BPSO-DT algorithm and more data the architecture had, the better accuracy the architecture could achieve.

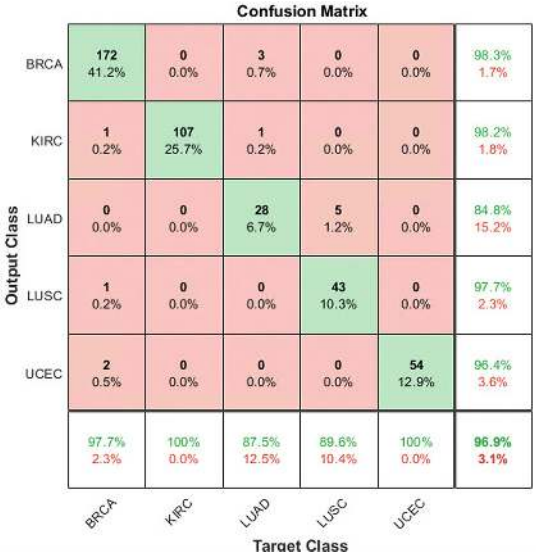
Another measure of performance is the progress of validation accuracy through the training phase. The progress of



Confusion Matrix

Output Class \ Target Class	BRCA	KIRC	LUAD	LUSC	UCEC	
BRCA	259 41.4%	1 0.2%	0 0.0%	1 0.2%	0 0.0%	99.2% 0.8%
KIRC	3 0.5%	158 25.2%	0 0.0%	0 0.0%	0 0.0%	98.1% 1.9%
LUAD	0 0.0%	0 0.0%	32 5.1%	4 0.6%	0 0.0%	88.9% 11.1%
LUSC	0 0.0%	2 0.3%	17 2.7%	66 10.5%	0 0.0%	77.6% 22.4%
UCEC	1 0.2%	0 0.0%	0 0.0%	1 0.2%	81 12.9%	97.6% 2.4%
	98.5% 1.5%	98.1% 1.9%	65.3% 34.7%	91.7% 8.3%	100% 0.0%	95.2% 4.8%

FIGURE 5. Confusion matrix for 70% training and 30% testing strategy.



Confusion Matrix

Output Class \ Target Class	BRCA	KIRC	LUAD	LUSC	UCEC	
BRCA	172 41.2%	0 0.0%	3 0.7%	0 0.0%	0 0.0%	98.3% 1.7%
KIRC	1 0.2%	107 25.7%	1 0.2%	0 0.0%	0 0.0%	98.2% 1.8%
LUAD	0 0.0%	0 0.0%	28 6.7%	5 1.2%	0 0.0%	84.8% 15.2%
LUSC	1 0.2%	0 0.0%	0 0.0%	43 10.3%	0 0.0%	97.7% 2.3%
UCEC	2 0.5%	0 0.0%	0 0.0%	0 0.0%	54 12.9%	96.4% 3.6%
	97.7% 2.3%	100% 0.0%	87.5% 12.5%	89.6% 10.4%	100% 0.0%	96.9% 3.1%

FIGURE 6. Confusion matrix for 80% training and 20% testing strategy.

validation accuracy shows the improvement of the leaning process. Figure 7 presents the progress of the validation accuracy through the training process while Figure 8 illustrates samples of testing classification accuracy using the proposed architecture.

B. PERFORMANCE EVALUATION AND DISCUSSION

To evaluate the performance of the proposed architecture, more performance measures need to be investigated through this research. The most common performance measures in the field of deep learning are Precision, Recall, and F1 Score [47] and are presented in equation (7), equation (8) and equation (9).

$$Precision = \frac{TP}{(TP + FP)} \tag{7}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{8}$$

$$F1Score = 2 * \frac{Precision * Recall}{(Precision + Recall)} \tag{9}$$

where The TP is an acronym for "True Positive" which represents the outcome where the model correctly predicts the positive class. Accordingly, The TN is an acronym for "True Negative" which represents the outcome where the model correctly predicts the negative class. Moreover, The FP is an abbreviation for "False Positive" which presents the outcome where the model incorrectly predicts the positive class. In addition, FN is the abbreviation for "False Negative" which presents the outcome where the model incorrectly predicts the negative class.

TABLE 3. Precision, Recall and F1 Score values for the different adopted training and testing strategies.

Training and Testing Strategy	Strategy 1 (50% -50%)	Strategy 2 (60% -40%)	Strategy 3 (70% -30%)	Strategy 4 (80% -20%)
Precision	89.85%	91.45%	90.72%	94.96%
Recall	90.15%	91.11%	92.30%	95.09%
F1 Score	90.69%	91.28%	91.50%	95.03%

Table 3 represents the values of Precision, Recall and F1 Score for the different adopted training and testing strategies. From the values presented in Table 3, the 80% training and 20% testing strategies give the best values which reflect the amount of data to be trained, the validation accuracy have been used through the training phase after every epoch as presented in Figure 7. The black circles showed up in Figure 7 presented the validation accuracy after every epoch of training. Using the augmentation techniques helped in generating more data which lead to a significant improvement in Precision, Recall, and F1 Score values.

Another measure of performance is the progress of validation accuracy through the training phase. The progress of validation accuracy shows the improvement of the leaning process. Figure 7 presents the progress of the validation accuracy through the training process while Figure 8 illustrates samples of testing classification accuracy using the proposed architecture.

The progress of validation accuracy has been improved through the training phase as after every epoch, the validation accuracy has been calculated and presented in Figure 7 with black circles while blue lines are the training accuracy.

The work presented in this research is novel in terms of pre-processing phase which include optimization process and the design of deep learning architecture. One of the related works presented in [36] made a similar contribution to ours. Table 4 illustrates the main difference between work presented in [36] and our presented work. The question that may raise itself in this section why compare our work with work presented in [36] and why not compared with the other related work in [5], [35], [38], [39].The main reason is the work presented in [36] uses the same methodology similar to ours

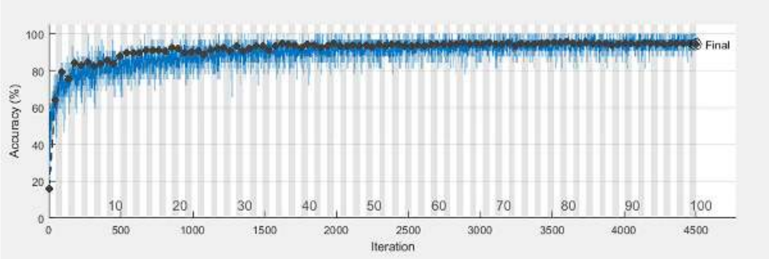


FIGURE 7. The progress of validation accuracy through the training phase.

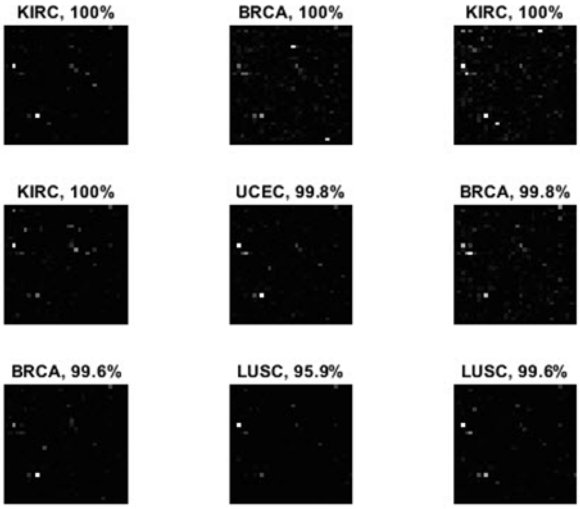


FIGURE 8. Samples of testing classification accuracy using the proposed architecture.

TABLE 4. A comparison between related work and the presented work.

	Paper [36])	This work
Dataset	Cancer Genome Atlas [48]	Tumor Gene Expression [40]
Number of Samples	10267 samples * 20531 Gene	2068 samples * 971 Gene
Optimization	no	Yes (BPSO-DT)
Pre-processing	y=log2(x+1) Variance threshold 102 * 102 image	BPSO-DT equation (1) no threshold 25 * 25 image
Deep learning Architecture	23 layers	14 layers
Number of Neurons in fully connected layers	36864 neurons in fc1 1024 neurons in fc2 33 neurons in fc3	64 neurons in fc1 5 neurons in fc2
Accuracy for BRCA	99.00%	98.30%
Accuracy for KIRC	95.00%	98.20%
Accuracy for LUAD	95.00%	84.8%
Accuracy for LUSC	91.00%	97.7%
Accuracy for UCEC	96.00%	96.40%
Accuracy for 5 classes	95.20%	96.90%

which included converting the RNA sequence to images then applying deep learning models, while the other related works used different methodologies which will be unfair to compare our work with them. Also, the selected dataset [40] which is used in this research in newly published in May 2018, it is an open opportunity to experiment the proposed model on this newly published data. Table 4 shows clearly that the proposed architecture in this work had lower training data but using the adopted augmentation techniques gave a better overall testing accuracy with 96.16% beating the related work in terms of overall testing accuracy for 5 classes. Moreover, this work did not apply any type thresholding the values and consider the training of the whole dataset, while the related work had dropped some of thresholding techniques on data. Additionally, the proposed deep learning architecture is power in complexity, as it had only 14 layers. whereas the related work had 23 layers will a huge number of neurons in fully connected layers with 36864 neurons in FC layer number 1 and 1024 FC layer number 2 which will reflect on the time of training on hardware. On the other hand, the related work achieved better testing accuracies for BRCA and LUAD, while ours achieved better testing accuracies for KIRC, LUSC and UCEC. Moreover, the related work was able to identify 33 classes, while this work identifies 5 classes.

VII. CONCLUSION AND FUTURE WORKS

Cancer is a group of diseases exhibiting abnormal cell growth, which may have the ability to invade or spread to various parts of the human body. This type of disease is responsible for more than 9 million deaths globally. The existence of RNA-Seq currently has greatly boosted the analysis of human genomics due improvements in the efficiency and accuracy, which help in understanding the nature of the cancer diseases. This paper introduced a novel approach to classifying different type of cancer: breast invasive carcinoma (BRCA), kidney renal clear well carcinoma (KIRC), E:at Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC) and Uterine Corpus Endometrial Carcinoma (UCEC). The proposed approach consisted of three phases. The first phase is the pre-processing, which included the optimization process using binary particle swarm optimization with design trees (BPSO-DT) algorithm to select the best features of RNA-Seq then converted it to 2D images. The second phase is the data augmentation, which increased the original dataset volume to 5 times larger. The third phase is the deep convolutional neural network architecture, in this phase, an architecture of two convolutional layers for feature extraction and two fully connected layers was presented to classify the 5 different cancer types. The presented results and the performance metrics preformed in this research showed that the proposed approach achieved an overall testing accuracy of 96.90%. The comparative results were introduced, and the accuracy achieved in the present work outperforms those of other related work for the testing accuracy for 5 classes of the tumour. Moreover, the proposed approach is less complexity and had less time in training. One of the potential future works is applying new architectures of deep neural networks such as Generative Adversarial Neural networks. GAN will be used before the proposed architecture. It will help in generating new images from the trained images, which will reflect on the accuracy of the proposed architecture. Additionally, to expand the current work to classify the 33 types of cancer if the datasets would be available with different deep learning architecture such as AlexNet, Vgg-16, and google-net.

ACKNOWLEDGMENT

The authors would like to thank NVIDIA Corporation with the donation of the Titan X GPU used for this research.

REFERENCES

[1] Y. H. Zhang, "Identifying and analyzing different cancer subtypes using RNA-seq data of blood platelets," *Oncotarget*, vol. 8, no. 50, pp. 87494–87511, Oct. 2017.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, A Cancer J. Clinicians*, vol. 68, no. 6, pp. 394–424, Nov. 2018.

[3] C. Hutter and J. C. Zenklusen, "The cancer genome atlas: Creating lasting value beyond its data," *Cell*, vol. 173, no. 2, pp. 283–285, Apr. 2018.

[4] F. Sanchez-Vega, "Oncogenic signaling pathways in the cancer genome atlas," *Cell*, vol. 173, no. 2, pp. 321.e10–337.e10, 2018.

[5] Y. Li, "A comprehensive genomic pan-cancer classification using the cancer genome Atlas gene expression data," *BMC Genomics*, vol. 18, no. 1, p. 508, Jul. 2017.

[6] B. Chopard and M. Tomassini, "Particle swarm optimization," in *An Introduction to Metaheuristics for Optimization*. Cham, Switzerland: Springer, 2018, pp. 97–102.

[7] M. N. Elbedwehy, H. M. Zawbaa, N. Ghali, and A. E. Hassanien, "Detection of heart disease using binary particle swarm optimization," in *Proc. Federated Conf. Comput. Sci. Inf. Syst. (FedCSIS)*, 2012, pp. 177–182.

[8] B. Gupta, A. Rawat, A. Jain, A. Arora, and N. Dhami, "Analysis of various decision tree algorithms for classification in data mining," *Int. J. Comput. Appl.*, vol. 163, no. 8, pp. 15–19, Apr. 2017.

[9] D. Rong, L. Xie, and Y. Ying, "Computer vision detection of foreign objects in walnuts using deep learning," *Comput. Electron. Agricult.*, vol. 162, pp. 1001–1010, Jul. 2019.

[10] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey," *Neurocomputing*, vol. 300, pp. 17–33, Jul. 2018.

[11] J. Maitre, K. Bouchard, and L. P. Badard, "Mineral grains recognition using computer vision and machine learning," *Comput. Geosci.*, vol. 130, pp. 84–93, Sep. 2019.

[12] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, May 2019.

[13] A. Maier, C. Syben, T. Lasser, and C. Riess, "A gentle introduction to deep learning in medical image processing," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 86–101, May 2019.

[14] J. Zhang, Y. Xie, Q. Wu, and Y. Xia, "Medical image classification using synergic deep learning," *Med. Image Anal.*, vol. 54, pp. 10–19, May 2019.

[15] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[16] E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang,Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu, T. Whyntie, P. Nachev, M. Modat, D. C. Barratt, S. Ourselin, M. J. Cardoso, and T. Vercauteren, "NiftyNet: A deep-learning platform for medical imaging," *Comput. Methods Programs Biomed.*, vol. 158, pp. 113–122, May 2018.

[17] S. Liu, Y. Wang, X. Yang, B. Lei, L. Liu, S. X. Li, D. Ni, and T. Wang, "Deep learning in medical ultrasound analysis: A review," *Engineering*, vol. 5, no. 2, pp. 261–275, Apr. 2019.

[18] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[19] G. Eraslan, . Avsec, J. Gagneur, and F. J. Theis, "Deep learning: New computational modelling techniques for genomics," *Nature Rev. Genet.*, vol. 20, no. 7, pp. 389–403, Jul. 2019.

[20] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, Feb. 2018, Art. no. 7068349.

[21] J. Riordon, D. Sovilj, S. Sanner, D. Sinton, and E. W. K. Young, "Deep learning with microfluidics for biotechnology," *Trends Biotechnol.*, vol. 37, no. 3, pp. 310–324, 2019.

[22] J. You, R. D. Mcleod, and P. Hu, "Predicting drug-target interaction network using deep learning model," *Comput. Biol. Chem.*, vol. 80, pp. 90–101, Jun. 2019.

[23] K. Jaganathan, S. Kyriazopoulou Panagiotopoulou, J. F. Mcrae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. Cui, G. B. Schwartz, E. D. Chow, E. Kanterakis, H. Gao, A. Kia, S. Batzoglou, S. J. Sanders, and K. K.-H. Farh, "Predicting splicing from primary sequence with deep learning," *Cell*, vol. 176, no. 3, pp. 535.e24–548.e24, Jan. 2019.

[24] C. Cao, F. Liu, H. Tan, D. Song, W. Shu, W. Li, Y. Zhou, X. Bo, and Z. Xie, "Deep learning and its applications in biomedicine," *Genomics, Proteomics Bioinf.*, vol. 16, no. 1, pp. 17–32, Feb. 2018.

[25] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3642–3649.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[28] D. C. Cire an, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Medical Image Computing and Computer-Assisted Intervention*. Berlin, Germany: Springer, 2013, pp. 411–418.

[29] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2015, pp. 730–734.

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[32] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[34] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[35] Y.-H. Hsu and D. Si, "Cancer type prediction and classification based on RNA-sequencing data," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 5374–5377.

[36] B. Lyu and A. Haque, "Deep learning based tumor type classification using gene expression data," in *Proc. ACM Int. Conf. Bioinf., Comput. Biol., Health Informat. (BCB)*, 2018, pp. 89–96.

[37] P. Danaee, R. Ghaeini, and D. A. Hendrix, "A deep learning approach for cancer detection and relevant gene identification," in *Proc. Pacific Symp. Biocomputing.*, vol. 22, 2016, pp. 219–229,

[38] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data," *Comput. Methods Programs Biomed.*, vol. 166, pp. 99–105, Nov. 2018.

[39] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "A deep learning-based multi-model ensemble method for cancer prediction," *Comput. Methods Programs Biomed.*, vol. 153, pp. 1–9, Jan. 2018.

[40] K. N. C. Ferles and Y. Papanikolaou, "Cancer types: RNA sequencing values from tumor samples/tissues," 2018. Distributed by Mendeley. [Online]. Available: https://data.mendeley.com/datasets/sf5n64hydt/1

[41] S. M. Vieira, L. F. Mendonça, G. J. Farinha, and J. M. Sousa, "Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients," *Appl. Soft Comput.*, vol. 13, no. 8, pp. 3494–3504, Aug. 2013.

[42] N. E. M. Khalifa, M. H. N. Taha, and A. E. Hassanien, "Aquarium family fish species identification system using deep neural networks," in *Proc. Int. Conf. Adv. Intell. Syst. Inform.*, 2019, pp. 347–356.

[43] N. E. Khalifa, M. Hamed Taha, A. E. Hassanien, and I. Selim, "Deep galaxy V2: Robust deep convolutional neural networks for galaxy morphology classifications," in *Proc. 2018 Int. Conf. Comput. Sci. Eng. (ICCSE)*, Mar. 2018, pp. 1–6.

[44] N. E. M. Khalifa, M. H. N. Taha, A. E. Hassanien, and I. M. Selim, "Deep galaxy: Classification of galaxies based on deep convolutional neural networks," Sep. 2017, *arXiv:1709.02245*. [Online]. Available: https://arxiv.org/abs/1709.02245

[45] N. Khalifa, M. Taha, A. Hassanien, and H. Mohamed, "Deep iris: Deep learning for gender classification through iris patterns," *Acta Inf. Med.*, vol. 27, no. 2, p. 96, 2019.

[46] A. A. Hemedan, A. E. Hassanien, M. H. N. Taha, and N. E. M. Khalifa, "Deep bacteria: Robust deep learning data augmentation design for limited bacterial colony dataset," *Int. J. Reasoning-Based Intell. Syst.*, vol. 11, no. 3, p. 256, 2019.

[47] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," in *Advances in Information Retrieval*. Berlin, Germany: Springer, 2005, pp. 345–359.

[48] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, "The cancer genome atlas pan-cancer analysis project," *Nature Genet.*, vol. 45, no. 10, pp. 1113–1120, Oct. 2013.

NOUR ELDEEN M. KHALIFA received the B.Sc., M.Sc., and Ph.D. degrees from the Information Technology Department, Faculty of Computers and Artificial Intelligence, Cairo University, Cairo, Egypt, in 2006, 2009, and 2013, respectively, and the M.Sc. degree in cloud computing from Cairo University, in 2018. He is currently an Assistant Professor with the Faculty of Computers and Artificial Intelligence, Cairo University. His research interests include wireless sensor networks, cryptography, multimedia, network security, and machine and deep learning. He is a Reviewer of the *Egyptian Informatics Journal*, *Ecological Informatics*, and *Internet of Things Journal* published by Elsevier.

MOHAMED HAMED N. TAHA received the B.Sc., M.Sc., and Ph.D. degrees from the Faculty of Computers and Artificial Intelligence, Cairo University, in 2006, 2009, and 2013, respectively. He has been an Assistant Professor with the Information Technology Department, Faculty of Computers and Artificial Intelligence, Cairo University, since 2016. He is a Reviewer of the IEEE INTERNET OF THINGS JOURNAL. Deep learning, machine learning, the Internet of Things, wireless sensor networks, and blockchain are his research interests.

DALIA EZZAT ALI received the B.Sc. degree in information technology from the Faculty of Computers and Artificial intelligence, Cairo University, Egypt, where she is currently pursuing the M.Sc. degree in information technology. Her current research interests are in the areas of deep learning, machine learning, and intelligent optimization.

ADAM SLOWIK (Senior Member, IEEE) was born in Warsaw, Poland, in 1977. He received the B.Sc. and M.Sc. degrees in computer engineering from the Department of Electronics and Computer Science, Koszalin University of Technology, Poland, in August 2001, the Ph.D. degree (Hons.) in electronics from the Department of Electronics and Computer Science, Koszalin University of Technology, in March 2007, and the Dr. Habilitation degree (D.Sc.) in computer science from the Department of Mechanical Engineering and Computer Science, Czestochowa University of Technology, Poland. Since October 2013, he has been an Associate Professor with the Department of Electronics and Computer Science, Koszalin University of Technology. He is the author or co-author of over seventy articles, and two books (in Polish). His research interests include soft computing, computational intelligence, machine learning, and bio-inspired global optimization algorithms and their engineering applications. Also, he is a member of the Program Committee of several International Conferences in the area of artificial intelligence and evolutionary computation. He is also an Associate Editor of the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, and a reviewer for many international scientific journals.

ABOUL ELLA HASSANIEN is currently the Founder and the Head of the Egyptian Scientific Research Group (SRGE) and a Professor of information technology with the Faculty of Computer and Information, Cairo University. He is also the Ex-Dean of the Faculty of Computers and Information, Beni Suef University. He has more than 1000 scientific research articles published in prestigious international journals and over 45 books covering, such diverse topics as data mining, medical images, intelligent systems, social networks, and smart environment. He is also a member of the Specialized Scientific Council of Information and Communications Technology, and Academy of Scientific Research and Technology (ASRT).