

## Deep Learning Approach for Identifying and Classifying Cancer Types Using BPSO and GAN

J.C. Prasad<sup>1</sup>, Megha Justin<sup>2</sup>, V.V. Revathy<sup>3</sup>, Sona Jose<sup>4</sup>, M. Sona Saju<sup>5</sup>

### Abstract

One of the most dangerous and dreadful diseases in the world is cancer. In 2020, it was found that over 10 million people will die due to cancer. Main causes of cancer are either due to inherited genetics or environmental and lifestyle factors. The risk of cancer increases significantly with age, their type and many cancers occur more commonly in developed countries. Early detection of cancer plays an important role, as it increases the possibilities of betterment. Latest improvement of artificial intelligence techniques in terms of productivity and precision and also optimization algorithms have largely smoothed out the human genomics study. So, we propose an RNA-Sequence analysis method for identifying and classifying cancer types, which uses Binary Particle Swarm Optimization using Support Vector Machine (BPSO-SVM) as feature selection tool and Generative Adversarial Network (GAN) along with different classical augmentation techniques to avoid overfitting by increasing the size of dataset and a deep learning network as the last phase for classification and thus helps in increasing accuracy.

**Keywords:** *Cancer, identification, classification, optimization, binary particle swarm optimization, support vector machine, augmentation, generative adversarial network, deep learning, convolutional neural network.*

### Introduction

The term 'Cancer' is used to represent a group of diseases, with irregular cell development, having intrusive properties and spreading to other parts of the body. Based on the data obtained from the study conducted by WHO, more than 8 million people die from cancer, accounting for about 13% of deaths worldwide, every year, showing cancer as one of the life threatening diseases in the universe.

---

<sup>1</sup> Department of Computer Science and Engineering, Federal Institute of Science and Technology, Kerala, India.  
E-mail: cheeran.prasad@gmail.com

<sup>2</sup> Department of Computer Science and Engineering, Federal Institute of Science and Technology, Kerala, India.  
E-mail: meghzj3888@gmail.com

<sup>3</sup> Department of Computer Science and Engineering, Federal Institute of Science and Technology, Kerala, India.  
E-mail: revathyvenugopal1998@gmail.com

<sup>4</sup> Department of Computer Science and Engineering, Federal Institute of Science and Technology, Kerala, India.  
E-mail: sonajose084@gmail.com

<sup>5</sup> Department of Computer Science and Engineering, Federal Institute of Science and Technology, Kerala, India.  
E-mail: sonasaju313@gmail.com

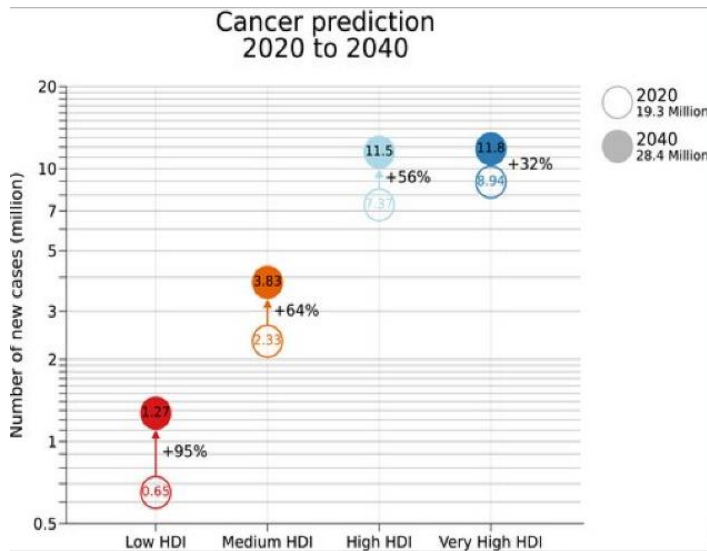


Fig. 1. According to the 4-Tier Human Development Index (HDI), the figure shows the projected number of new cases for all cancer types of both sexes combined in 2040.

Cancer stands as a leading root for death and an important block for increasing life expectancy in every country. Around 17% of females and 20% of males will be affected by cancer at some point, out of which 10% of females and 13% of males will die. According to estimates from WHO, in 2019, cancer is considered as the first or second leading cause for death prior to the age of 70 years out of 112 in 183 countries and ranks third or fourth in other 23 countries. Cancer's rising prominence as one of the main reasons for death partly shows marked declines in death rates of stroke and coronary heart condition, relative to cancer, in many countries.

RNA-Sequence, which was developed in the mid of 2000's has the potential to profile biomarkers used for clinical indications, new disease biology identification, inferring druggable pathways, and to make genetic diagnoses. These obtained results could be further personalized for several subgroups or for patients individually, thus potentially highlighting more effective prevention methods, diagnostics, and therapy. RNA-Seq technology can detect a high percent of genes with low expression, ie, differentially expressed genes and thereby increasing the specificity and sensitivity in disease analysis. Hence, RNA-Sequence data stands better than image data.

The Cancer Genome Atlas (TCGA) is a breakthrough in the genomics cancer dataset, which includes DNA methylation, variations in DNA sequence of somatic cells with exclusive reproductive cells, structural changes, expression of proteins, gene and microRNA. Between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) under TCGA, a combined project was entrenched that served as a model project in 2006 and focussed on 3 cancer types: ovarian, glioblastoma and lung. NHGRI and NCI in 2009, relicensed TCGA for a finished manufacturing phase. TCGA collected more than 11,000 cases of 33 types of tumours and produced a huge, encyclopedic database defining molecular variations that appear in tumours, in the upcoming years, thus providing an appreciable classification opportunity of divergence at DNA, protein levels and RNA for the universal outlook.

Most of the existing models have advantages as well as disadvantages. This paper proposes a model to distinguish normal and tumour samples that depends on a high-D RNA-Sequence data, which will increase the accuracy in predicting by selecting the best features out of the total features available in

the dataset for each cancer type using Binary Particle Swarm Optimization using Support Vector Machine (BPSO-SVM). Also, to avoid overfitting problem which occurs due to less number of samples available with respect to the learnable parameters of deep learning phase that is, by increasing the size of dataset using classical augmentation techniques like rotation, flipping, adding noise along with it implementing Generative Adversarial Network (GAN), which generates images from trained images, thereby increases number of samples of cancer types. This helps in classifying those types which have only a very limited number of samples available, because of which they are usually excluded from classification.

### Literature Review

In the paper, “The efficacy of the various machine learning models for multi-class classification of RNA-seq expression data”[1], Random Forest (RF), along with other ensemble machine learning algorithms like, Gradient Boosting Machine (GBM), and Random Ferns (RFERN) are the methods used. The performance of these algorithms were compared with a clustering and a classification algorithm; K-Nearest Neighbor (KNN) and Support Vector Machine(SVM), respectively. The random forest algorithm has diagnosed 17 different types of cancer with overall accuracies of 99.89% and the gradient boosting machine attained an accuracy of 99.68%, respectively.

In “A deep learning approach for cancer detection and relevant gene identification”[2], as the first step, features were extracted from high dimensional gene expression profiles using Stacked Denoise Autoencoder (SDAE). In the next step, Performance evaluation is done on the extracted features. Finally, a set of highly interactive genes were identified. Limitation of this deep learning approach is there is a need for big data sets. It achieved an accuracy of 94.78%.

In “Deep Learning Based Tumor Type Classification Using Gene Expression Data”[3], preprocessing is done in the beginning on the input data samples. Then the genes with small variance are filtered out from the total samples. The high-dimension expression data is then embedded into a 2-D image, which is then fed for classification, which is done using a convolutional neural network. As the third step, heat map generation is done for each class and genes with top intensities are selected. In the fourth step, validation of the pathways of selected genes is carried out. Accuracy obtained using this model is 95.59%, which is higher than other papers using GA/KNN method on the same dataset. On the genes with top strengths in heat-map, functional analysis is applied and top genes which are validated are related to tumor-specific pathways, and have been already used as biomarkers, which proves the effectiveness of this method. This is the first paper to implement convolutional neural networks on Pan-Cancer Atlas for tumour type classification, and also to match the significance of classification with the importance of genes. Experiment results show that this method has a good performance and can also be applied in other genomics data.

In the paper, “A deep learning based multi-model ensemble method for cancer prediction”[4], the first step is the feature selection using the DESeq method. In the second step, S-fold cross validation is used, to reduce the generalization error and prevent over-fitting. As the third step, output of the second step is given to the first stage classification model and as the fourth step, it is then given to the second stage classification model. The advantages of each classifier are fully considered and utilized. The deep learning-based multi-model ensemble method reduces the generation error and obtains more information by using the first-stage predictions as features.

The method used for cancer prediction in, “A semi- supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data”[5], is Stacked Sparse Auto-Encoder

(SSAE) based semi-supervised classification model. In the first step, gene expression data is being analyzed. In the next step, the greedy layer-wise pre-training strategy is applied so as to obtain a layer-by-layer initialization. Finally, the Supervised neural network classifier is linked to the middle layer of the SSAE. One of the limitations for this model is that the info given by the finite labeled data may not be enough for model prediction. It achieved an accuracy of 98.15%, 96.23%, and 99.89% for the STAD, BRCA and LUAD datasets.

In the paper, “Lightweight convolutional neural network for breast cancer classification using RNA-seq gene expression Data” [6], at first, Gene expression data is preprocessed by removing the outlier samples by keeping a cut-off of 0.6, decided depending on the Array-Array Intensity Correlation (AAIC) that determines the square matrix of spearman and then the data gets transformed into 2D-images. After which normalization is done using the tcga normalizing function so as to ensure that the expression levels inferred from it are correct and to keep away predisposition in expression estimation. Finally, filtering is done on the data which is then fed into a convolutional neural network. To overcome the problem of small size and high dimensionality, 2 convolutional layers are used in which hyperparameters are selected carefully i.e., only hyperparameters with high values are selected using the grid search approach with 5-fold cross validation. There are several hidden points due to overlapping and also there may be data loss as the outliers are removed. It achieved an overall accuracy of 98.76%.

A new method Genome Deep Learning (GDL) involving Deep neural network (DNN) model is used in “Identification of 12 cancer types through genome deep learning” [7], for cancer identification which is based on genomic changes. Its architecture consists of feature selection, feature quantization and data filters. Deep neural networks have several multiple hidden layers between input-output layers. GDL has two methods, namely, data processing and model training. The overall accuracy, sensitivity and specificity of the total-specific model were found to be 94.70%, 97.30% and 85.54%, respectively for cancer identification.

“Diagnosis of breast cancer using a combination of genetic algorithm and artificial neural network in medical infrared thermal imaging” [8], uses the combinatorial model that has an Artificial Neural Network (ANN) and Genetic Algorithm (GA). All the diagnostic specifications are inserted into the combinatorial model, and the system, by analyzing the information, will then select and extract the best diagnostic parameters. The results of the combinatorial model with 50% sensitivity, 75% specificity and 70% accuracy display proper precision in cancer diagnosis. One of the limitations is that it hasn't found an accurate model for the classification of every cancer pattern.

“A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data”[9], uses Genetic Algorithm (GA) and k-nearest neighbours (KNN) for identification and classification of cancer. In this, GA is used as the feature selection engine and KNN as classification tool. Genetic algorithms encode chromosomes into gene signatures. This method can thus identify gene signatures that not only can separate different classes but also their subtypes. Identified almost 31 tumor types with over 90% of correctness from the whole samples.

To detect lung cancer at its early stage, “Lung cancer detection by using artificial neural network and fuzzy clustering methods” [10], model uses 2 segmentation methods, namely, Hopfield Neural Network (HNN) and a Fuzzy C-Mean (FCM) clustering algorithm, for segmenting sputum color images. The Threshold algorithm has thrived in extracting the nuclei and cytoplasm regions. Moreover, it prospered in finding out the best range of thresholding values. Major drawback is that Fuzzy C-Mean (FCM) failed

in detecting the nuclei, instead it detected only a part of it. Thresholding classifiers have obtained a good accuracy of 98% with significant values of sensitivity and specificity of 83% and 99% respectively.

### A. Binary Particle Swarm Optimization

Particle swarm optimization (PSO) is a good performance and optimization technique made by Kennedy and Eberhart. It is a population-based search, which is based on the organism's perspective on a social setting, of which a bird flock or a fish school are examples. It does not need high computer speeds and therefore includes only simple statistics. On top of all that, it only uses a small number of parameters to perform the adjustment and the same parameters are used for various alternatives to ongoing-use problems and problems of incomprehensible performance such as feature selection problems. In PSO, the whole solution of the problem is considered as a particle and this group of particles called a swarm lies in the D-dimensional search space, so as to find sound solutions by validating the position of each particle according to its experience as well as its neighbouring particles. The parameters of are:

- *Swarm size*: The number of particles in the swarm.
- *Neighbourhood size*: Defines the extent of social interaction within the swarm.
- *Number of iterations*: The number of iterations to reach a good solution is also problem-dependent.
- *Velocity*: The mechanism used to move (evolve) the position of a particle to search for optimal solutions.
- *Acceleration coefficients*:  $c_1$  and  $c_2$ , the acceleration coefficients, which together with random vectors  $r_1$  and  $r_2$ , control the stochastic influence of the cognitive and social components on the overall velocity of a particle.
- *Inertia weight*: It determines the characteristic of maintaining the particle's inertial motion and expansion of the search space.
- *Local best(Personal best(pbset))*: Best solution obtained by a particle obtained so far.
- *Global best(gbest)*: Best value achieved by any particle so far by any particle in the neighbourhood of that particle.

All particles have values of fitness, found using a function, and the speeds that will lead to the motion of particles. The particle's position,  $i$  at  $k$  iteration during its motion, is represented by a vector,  $X_i^k = (x_i^1, x_i^2, \dots, x_i^D)$  and the velocity of particle  $i$  at  $k$  iteration is denoted as  $V_i^k = (v_i^1, v_i^2, \dots, v_i^D)$ . Each particle validates its velocity and position based on the local fitness value (Pbest) and the global fitness value (Gbest). so far by any particle in the neighborhood of that particle.

$$\begin{aligned} V_i^k &= wV_i^{k-1} + c_1r_1(Pbest_i - X_i^{k-1}) + c_2r_2(Gbest - X_i^{k-1}) \quad (1) \\ X_i^k &= X_i^{k-1} + V_i^k \quad (2) \end{aligned}$$

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. Kennedy and Eberhart extended PSO to binary PSO (BPSO) [11], which can be used to solve discrete problems. Here, Pbest<sub>*i*</sub> and Gbest are made finite so, can take only values 1 or 0, and thus, the position update equation becomes a probabilistic equation. A sigmoid function sig( $V_i^k$ ) is used to convert the  $V_i^k$  to the range of (0,1).

$$\text{sig}(V_i^k) = 1 / (1 + e^{-k_i}) \quad (3)$$

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

Where, 1 implies that the feature is considered as an important one for the next generation and else not and the rand represents a random number [0, 1]. To optimize the high-dimensional RNA-Sequence to select only the most important features from the dataset, Binary Particle Swarm Optimization (BPSO) with Support Vector Machine(SVM) as its classifier can be used. SVM has been successfully applied to gene expression data classification problems, since they are not negatively affected by high dimensionality, hence they can obtain a higher accuracy than other classification methods [12]. This step helps in feature reduction, which in turn leads to the need for fewer resources to complete the computations. Less computation time and less storage capacity needed means the computer can do more work. Feature reduction also removes multicollinearity, resulting in improvement of the machine learning model.

### **B. Augmentation Using Classical Augmentation Techniques**

The Deep learning model has a huge number of graspable variables compared to the number of images in the training set because of which there is a chance of overfitting, and also deep learning models perform better with large datasets. So, data augmentation is done which can increase the size of the dataset. The different augmentation techniques applied on images include rotating the image with respect to different angles, taking reflection with respect to different axes, flipping the image horizontally as well as vertically, adding noise to the image, cropping the image, increasing and decreasing the brightness of the image etc. This will make the image invariant to rotation, reflection and other transformations. This assists in building simpler and robust models which helps to generalize better. This leads to remarkable development in the phase of neural network training and to be more robust and accountable for the testing phase. Different augmentation techniques are applied on the dataset taken after the pre-processing phase, which contains 2-D images of each sample. This helps to enlarge the dataset by increasing the number of samples for each cancer type.

### **C. Generative Adversarial Network**

Labeled medical data is scarce and costly to generate. Large amounts of data are needed to obtain generalizable deep learning models. A novel method for data augmentation is offered by Generative adversarial networks [13]. It consists of two neural networks, a generator and a discriminator. In a discriminator, a convolutional neural network is used, where it is first trained with a stack of matrices and then it becomes a smaller matrix and at last it forms a vector. In the generator, it first has a vector and enlarges its dimension. Generator takes a random noise and generates a sample of data. Discriminator is trained with real data as well as the sample generated by the generator. The discriminator network decides whether the data is generated or taken from the real sample using a binary classification problem, with the help of a sigmoid function, that gives the output in the range 0 to 1. The generative model analyzes the distribution of the data in such a way that, after the training phase, the probability of the discriminator making a mistake maximizes. The discriminator, on the other hand, is based on a model that will estimate the probability that the sample is coming from the real data or not from the generator.

### **D. Deep Learning Phase**

Deep learning in the field of computer vision, has achieved a dramatic change, in particular, in image acquisition, segregation and recognition, and is considered a development automated diagnostic

systems help to achieve superior results, as well as performance increase the scope of the disease, and, in turn, make real-time medical thinking on disease classification models. Artificial Intelligence branch, depending on the various algorithms for data processing and simulation of rational process or conceptualization. DL uses layers of algorithms for processing, analysis and detection of hidden patterns in data and visual objects recognition. Info passes from each layer of a deep network, with output of the preceding layer providing input for the next layer. The input layer is the first layer, whereas, the output layer is the last layer and all the layers between are referred to as hidden layers of the network. Each layer is simple and algorithm-friendly which contains a specific type of activation function. Also, advances in deep Convolutional Neural Network (CNN) have helped to reduce the level of error.

### **E. Comparison Table**

Related works that are very much similar to the implemented model are taken for consideration and are compared by looking into the methods used, advantages and limitations. This helped in gaining better understanding of the methods that can be used for identifying and classifying cancer types, which indeed gave an overview about the existing approaches and how they can be improved. The implemented model used methods such as BPSO-SVM for feature extraction and GAN for augmentation, achieving 99.5% accuracy.

*Table 1*

Comparison Table

Paper title	Brief about methodology	Advantages	Shortcomings (if any)
<b>The efficacy of various the machine learning models for Multi-class classification of RNA-seq expression Data</b>	Random Forest (RF), Gradient Boosting Machine (GBM), and Random Ferns (RFERN).	The random forest algorithm has diagnosed 17 types of cancer with accuracies of 99.89% and the machine were capable of 99.68%, respectively.	
<b>A Deep Learning Approach For Cancer Detection And Relevant Gene Identification</b>	Stacked Denoising Auto Encoder (SDAE).	Highly interactive genes could be useful cancer biomarkers for the detection of breast cancer.	Requires large no.of data sets, which may not be available for cancer tissues.
<b>Deep Learning Based Tumor Type Classification Using Gene Expression Data</b>	Convolutional Neural Network(CNN).	The accuracy is 95.59%. It has a good performance and could also apply in other genomics data.	Imbalanced dataset.Some tumor type has over 1000 samples while some only have 30 samples.
<b>A deep learning-based multi-model ensemble method for cancer prediction</b>	Support Vector Machines (SVMs), K-Nearest Neighbours (KNN) and Naive Bayes (NBs).	The intricate relationships among the classifiers are learned automatically, thus enabling the ensemble method to achieve better prediction.	Incurs a higher computational cost.
<b>A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data</b>	Stacked Sparse Auto-Encoder (SSAE).	The SSAE based semi-supervised classification method outperforms other classification methods and results in more accurate and stable predictions.	Requires a large amount of computational time, mainly due to the pre-training process of deep network structure.



Paper title	Brief about methodology	Advantages	Shortcomings (If any)
<b>Lightweight Convolutional Neural Network for Breast Cancer Classification Using RNA-Seq Gene Expression Data</b>	Pre-processed and transformed into 2D-images after which normalization is done and filtering is applied on the data. Finally a Convolutional Neural Network(CNN) phase.	Achieves an accuracy of 98.76%.	There are hidden points due to overlapping and there may be loss of data since outliers are removed.
<b>Identification of 12 cancer types through genome deep learning</b>	Genome Deep Learning) involving DNN(Deep neural network).	The accuracy, sensitivity and specificity of the total-specific model were 94.70%, 97.30% and 85.54%, respectively for cancer identification.	More factors in addition to genomic variations (such as age, sex, transcriptome and proteome data) might be integrated into the model to promote prediction accuracy.
<b>Diagnosis of Breast Cancer using a Combination of Genetic Algorithm and Artificial Neural Network in Medical Infrared Thermal Imaging</b>	Combinatorial model consisting of Backpropagation Neural Network and Genetic Algorithm.	The results of the combinatorial model with 50% sensitivity, 75% specificity and 70% accuracy show proper precision in cancer diagnosis.	They haven't found an accurate model for the classification of every cancers' pattern.
<b>A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data</b>	Genetic Algorithm (GA) and the k-nearest neighbors (KNN).	Identified many sets of 20 genes that could correctly classify more than 90% of the samples from 31 different tumor types.	Cancer types with less number of samples are not considered.
<b>Lung Cancer Detection by Using Artificial Neural Network and Fuzzy Clustering Methods</b>	2 segmentation methods: Hopfield Neural Network (HNN) and a Fuzzy C-Mean (FCM) clustering algorithm.	The thresholding classifier has achieved a good accuracy of 98% with high value of sensitivity and specificity of 83%.	Fuzzy C-Mean (FCM) failed in detecting the nuclei, instead it detected only part of it.

### Methodology

#### A. Data Description

The dataset is taken from Kaggle, and contains gene expressions of patients having different types of cancer: Kidney Renal Clear Cell Carcinoma (KIRC), Breast Invasive Carcinoma (BRCA), Lung Squamous Cell Carcinoma (LUSC), Lung Adenocarcinoma (LUAD), Uterine Corpus Endometrial Carcinoma (UCEC) in a CSV format having a data file and labels file. The number of samples(instances), which are stored row-wise is 801. The number of variables (attributes) ie, features of each sample are RNA-Sequence gene expression levels measured by illumina Hi-Seq platform is 20531. These cancer types are encoded numerically as: 1 - BRCA, 2 - KIRC, 3 - LUAD, 4 - LUSC, 5 - UCEC. Out of 20531 features, only 800 are taken, since only limited computational power and time is available. The data is then normalized to lie in the range [-1,1], so as to make the data into a standard format for further processing.

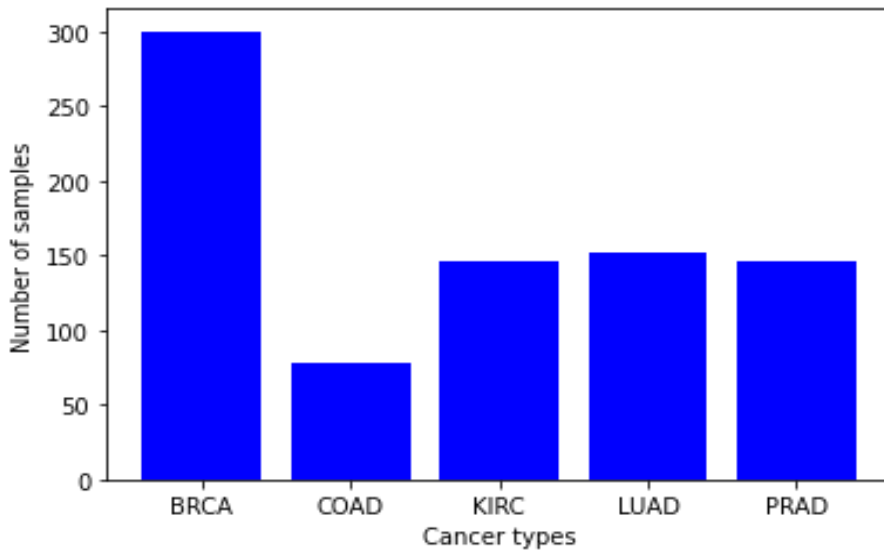


Fig. 2. Histogram Showing Number of Samples for Each Cancer Type

## B. Model Construction

### 1.) Pre-processing using Binary Particle Optimization using Support Vector Machine

BPSO is implemented to optimize the number of RNA-sequence features to minimum and select only important ones, and thus to increase the accuracy of the classification, and the Support Vector Machine (SVM) is taken as BPSO's fitness function in the equation (5) for the classification problem.

$$\text{Fitness} = \alpha(1 - C_p) + (1 - \alpha)(1 - (S_f / T_f)) \quad (5)$$

Where,  $\alpha$  is a hyperparameter that decides the trade-off between the classifier performance  $C_p$ , and the size of the feature subset  $S_f$  with respect to the total number of features  $T_f$ .

The steps of PSO are presented in Fig. 3.

**Input** : Tumor gene expression dataset

**Output**: Gbest position

```

1 Initialize the position and velocity of each particle
  randomly
2 while iteration condition is not satisfied do
3   Evaluate the fitness of the particle swarm by DT
   according to equation 5
4   for each particle i do
5     if the fitness of  $x_i$  is greater than the fitness of
     the  $Pbest_i$  then
6       |  $Pbest_i = x_i$ 
7     end
8     if the fitness of any particle of the swarm is
     greater than  $Gbest$  then
9       |  $Gbest = \text{particle's position}$ 
10    end
11    for each dimension  $D = 1, \dots, N$  do
12      | Update particles velocity and particles
      | position according to equation 1,3, and
      | 4 respectively
13    end
14  end
15  go to next generation until termination criterion is
   met
16 end
17 Output Gbest

```

Fig. 3. PSO Algorithm

The values given to different parameters are: number of iterations =10,  $c1 = c2 = 1.5$ ,  $w = 0.7$ ,  $v_{max} = 6$ ,  $v_{min} = -6$ , dimension= 800(Number of features), number of particles = 800, swarm size = 30. The output, ie, Gbest position of the particles are expressed in two terms: 1 or 0 (or on and off). Initially, each feature is assigned as the dimension of a particle. After implementing BPSO, we will obtain the best position, which can be interpreted as a binary array. The columns with value 0 are removed from the dataset, thereby reducing the features from 800 to 484.

The pre-processing phase also includes:

- Loading of the tumor gene expression on memory.
- Each row of the dataframe, which corresponds to the gene information of each of the samples is then converted to an image range.
- Constructing images by converting the optimized data into a  $22 \times 22$  pixels image and saving it.

Table 2

Number of each Cancer Type after BPSO-SVM

Cancer type	Number of samples
BRCA	300
COAD	78
KIRC	146
LUAD	141
PRAD	135

## 2.) Augmentation Using Classical Augmentation Techniques

The collection of images of the cancer types after the first phase is converted to an array form, so as to apply transformations. By applying the classical augmentation techniques, that makes the resulting model to pick images randomly from the existing dataset and apply transformations, such as:

- Rotation, (here, with respect to 25 degree)
- Horizontal flipping,
- Random noise

Over 20 new augmented images of each cancer type are created, which makes the cancer type with a smaller number of samples to be considered for the identification and classification process. This process will thus lead to improvement in the neural network training phase as it makes the dataset more invariant to noise, reflection and rotation.

## 3.) Generative Adversarial Network

The feature values from each image are extracted and then the image is converted into a numpy array and each of the pixel values are converted in between 0 and 1.

Discriminator consists of several layers. For down-sampling, first layer is taken as convolutional layer with 128 neurons, the LeakyReLU activation function is applied after which for classification, a flattening layer, a dropout rate of 0.4 is applied and finally a fully connected layer, which gives either 1 or 0 as output. The generator is an inverse convolutional network, i.e., it takes a vector of random noise and up-samples it to an image. For that, a convolutional layer is used after which the LeakyRelu activation function with  $\alpha=0.2$  is applied and then to reshape the array without changing data, the Reshape function is applied. After these steps, to upsample the image, Conv2D\_Transpose is used and then again, a LeakyRelu activation function is applied and then to generate an image a convolutional layer is used.

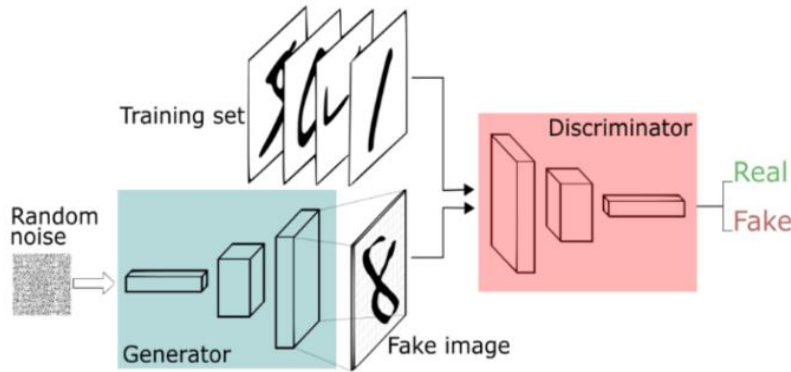
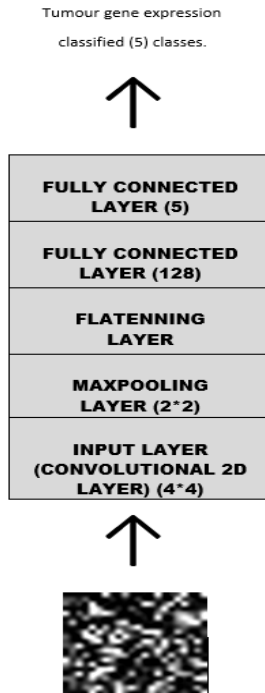


Fig. 4. Generative Adversarial Network

After individually defining both the CNNs for the generator and discriminator, they are combined by simply stacking the generator on top of the discriminator, so that the output of the generator is fed directly into the discriminator. Real samples are generated by taking the entire image collection array and randomly selecting a sample from it. By training the GAN, the weights of the discriminator and the generator are presumed to be linked, and the gradients are propagated backwards. To control the fitting of the model or computation time, the batch size can be adjusted. To find the overall performance, the performance of the discriminator is evaluated by just giving a fake image and asking the discriminator to predict the label. Then comparison is made between the predicted label with the original label and the accuracy is measured. Finally, the model is saved in.h5 format after several epochs.

The training process consists of giving a noisy vector as input into the generator, which generates the image and is sent to the discriminator, where it is asked to predict the output label. The discriminator checks the image dataset given to it and predicts the label. If the labels are mismatched, ie, since the original label for the noisy input is fake, while the discriminator says it's real, then there is an error. So the complete process of training would be to minimize the error and is done at each iteration by adjusting the weights of the CNNs. Hence, a fully trained GAN will be capable of generating real looking images from the noisy input. After loading the trained model file, weights will be re-initialized in a blank GAN network after which again some noisy vector input is fed into the generator, which will produce images similar to real images. The images are then saved. Number of epochs is fixed to 51, as both the loss function values of discriminator and generator are approximately the same at the end of 51 epochs and batch size is taken as 13, ie, for each epoch 13 random images are taken for the training process. Images are generated in such a way that each class i.e., each type of cancer has a total of 400 images (original + augmented + GAN generated), which in turn makes the dataset balanced.

#### 4.) Convolutional Neural Network



*Fig. 5.* Layer-wise Description of Proposed Deep Learning CNN Architecture for Classifying Cancer Type

A 2D CNN model is implemented. The dataset after the augmentation phase is taken and is converted to a suitable form for training. This is then split into 80% training and 20% testing sets. In the first layer, input layer, which is a convolutional of dimension (19,19,1) with (4,4) kernel padding and then the output of first layer is fed into max pooling layer of size (2,2) and dimension (9,9,1) and as the third layer, flattening layer is used to flatten the matrix into a single dimensional vector(1,2592). As the fourth layer, a fully connected layer with 128 input neurons and output shape is (1,128). Last layer i.e., output layer which is a fully connected layer with size (1,5) as there are 5 classes i.e., 5 different cancer types. In this layer, we use the softmax function which gives the output in the form of probabilities. The one with the highest probability will be declared as the output class.

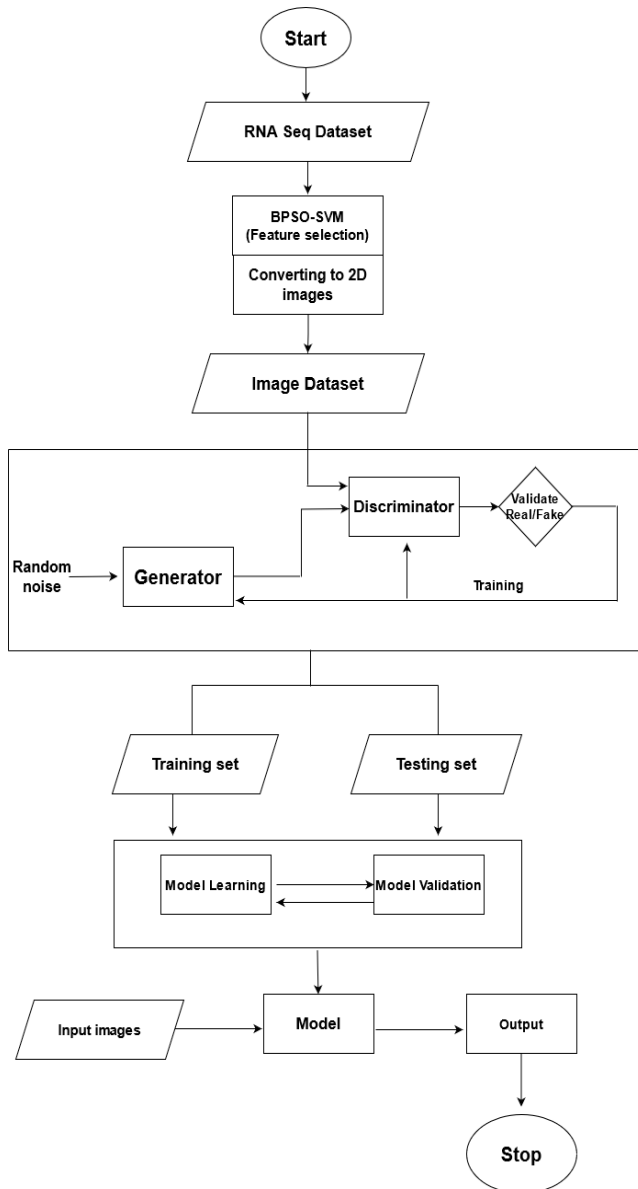


Fig. 6. Architecture

### C. Model Training

After the model is constructed, 80% of the dataset is given for training. A batch size, 50, is mentioned which takes into account the number of input samples at each epoch of training. The training is purely done on CPU using tensorflow. Validation accuracy of 99.5% is achieved.

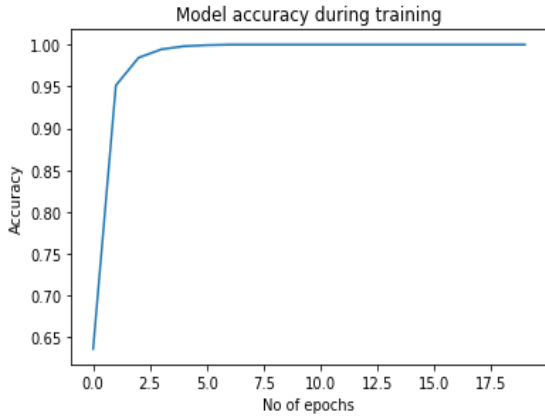


Fig. 7. Progress of Validation Accuracy during Training

#### D. Model Testing

To evaluate the performance of the proposed approach, several measures are used, including the Confusion Matrix, Precision, Recall, and F1-score. The metrics are defined by the confusion matrix. The evaluation parameters are verified by the inbuilt sklearn functions.

A confusion matrix for actual and predicted classes is derived from the standard five values namely TruePositive, FalsePositive, TrueNegative and FalseNegative to evaluate the performance.[14]

TP : TPos; TN : TNeg; FP : FPos; FN : FNeg

- 1) Accuracy: It is a good predictor for the degree of correctness in a training of the model and how it performs generally. It may be defined as the measure of the correct predictions in correspondence to the wrong ones.

$$Accuracy = \left( \frac{TPos + TNeg}{TPos + FPos + TNeg + FNeg} \right) \quad (6)$$

- 2) Precision: The degree of correctness in determining the positive outcomes. It is basically the ratio between true positives and the overall set of positives. This depicts the handling capacity of the system for positive values but does not provide insight into the negative values.

$$Precision = \left( \frac{TPos}{TPos + FPos} \right) \quad (7)$$

- 3) Recall: Also known as sensitivity, can be defined as the ratio of correctly determined positive instances to all observations. It is a measure for the effectiveness of the system in predicting positive outcomes and determining the costs.

$$Recall/Sensitivity = \left( \frac{TPos}{TPps + FNeg} \right) = \left( \frac{TPos}{Pos} \right) \quad (8)$$

- 4) F1-score: It is the weighted average of Recall and Precision. This measure considers both types of false values. F1 score is considered perfect when at 1 and is a total failure when at 0.

$$F1score = \left( \frac{2 \times (Precision \times Recall)}{Precision + Recall} \right) \quad (9)$$

The diagonal values represent the number of outputs correctly predicted for each class and others show the number of incorrectly predicted outputs. When normalization is done, the values are confined to lie in between the range [0,1].



Two confusion matrices are shown here, one without applying normalization and other by applying normalization.

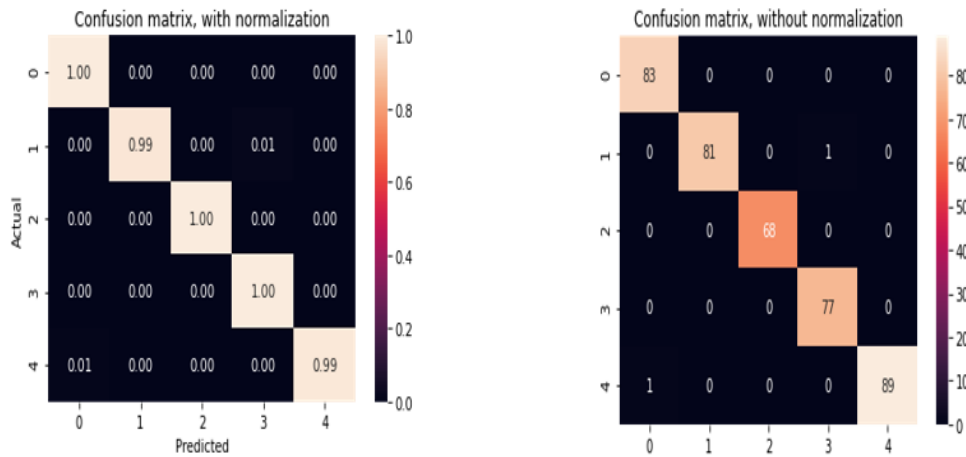


Fig. 8. Confusion Matrices

The overall accuracy obtained is 99.5%. Precision, Recall, F1-score values for each cancer type are shown below.

Table 3  
Performance Metrics

	BRCA	COAD	KIRC	LUAD	PRAD
Accuracy	0.995				
Precision	1	0.987	1	1	0.988
Recall	0.988	1	1	0.987	1
F1-score	0.99	0.99	1	0.99	0.99

### E. Model Evaluation

After the model is being trained and outputs are recorded, the model can be saved using the model. save function in Keras. This can be later loaded for testing purposes, by just reloading the model and then passing the test data into it.

Path of the folder, which contains images from 5 cancer types is given as input, from which an image is randomly selected and tested. The type of cancer is predicted as an output.

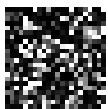
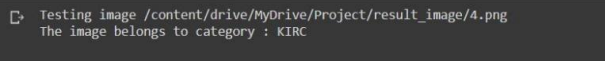


Fig. 9. Input



Testing image /content/drive/MyDrive/Project/result\_image/4.png  
The image belongs to category : KIRC

Fig. 10. Output

## Conclusion

The presence of RNA-seq largely enhanced the human genomics analysis because of the developments in the efficiency and accuracy, which helped in grasping the nature of the cancer diseases. Identifying cancer is of utmost importance as people diagnosed earlier with cancer are not only more likely to survive, but also have better experiences of care, lower treatment morbidity, and improved quality of life compared with those diagnosed late. Classifying cancer into its corresponding type correctly is central as it helps in developing knowledge, making diagnosis, assigning treatment and for specific drug development. There are several algorithms and methods available for feature selection like, Genetic Algorithm, DESeq method and so on. Also as classification tools different types of deep learning networks are being used.

The proposed model consists of Binary Particle Swarm Optimization using a Support Vector Machine (BPSO-SVM) in the pre-processing phase for feature selection and also to reduce the high dimensions of RNA-Seq, which is the input to this model. Classic augmentation techniques along with Generative Adversarial Network (GAN) are also included to avoid overfitting. And finally, as the last phase, a deep learning network for classification of cancer types. An overall accuracy of 99.5% is obtained.

In the future, the proposed method can be extended by classifying a wider variety of cancer types than the proposed model. Also, by the usage of larger computational resources more features can be taken into consideration, which further improves the model in correctly predicting the output.

## References

1. Padideh Danaee, Reza Ghaeini and David A. Hendrix, "A deep learning approach for cancer detection and relevant gene identification." *Pacific Symposium on Biocomputing*, 2017.
2. Boyu Lyu and Anamul Haque, "Deep learning based tumor type classification using gene expression data".
3. Yawen Xiaoa, Jun Wub, Zongli Linc and Xiaodong Zhao, "A Deep Learning based Multi-model Ensemble Method for Cancer Prediction".
4. Yawen Xiao, Jun Wu, Zongli Lin and Xiaodong Zhao, "A semi-supervised deep learning method based on a stacked sparse auto-encoder for cancer prediction using RNA-seq data".
5. Yuanyuan Li, Kai Kang, Juno M. Krahn, Nicole Croutwater, Kevin Lee, David M. Umbach and Leping Li, "A comprehensive genomic pan-cancer classification using the Cancer Genome Atlas gene expression data".
6. Murtada K. Elbashir, Mohamed Ezz, Mohanad Mohammed and Said S. Saloum, "Lightweight Convolutional Neural Network for Breast Cancer Classification Using RNA-Seq Gene Expression Data" *IEEE Access* (Volume:7), 2019.
7. Yingshuai Sun, Sitao Zhu, Kailong Ma, Weiqing Liu, Yao Yue, Gang Hu, Huifang Lu and Wenbin Chen, "Identification of 12 cancer types through genome deep learning", 2019.
8. Fatma Taher, Naoufel Werghi, Hussain Al-Ahmad and Rachid Sammouda, "Lung Cancer Detection by Using Artificial Neural Network and Fuzzy Clustering Methods", *American Journal of Biomedical Engineering*, 2012.
9. Sterling Ramroach, Melford John and Ajay Joshi, "The efficacy of various machine learning

- models for multiclass classification of RNA-seq expression data”, *Springer’s Intelligent Computing - Proceedings of the Computing Conference*, 2019.
10. Hossein Ghayoumi Zadeh, “Diagnosis of Breast Cancer using a Combination of Genetic Algorithm and Artificial Neural Network in Medical Infrared Thermal Imaging”, *Iranian Journal of Medical Physics*, 2012.
  11. Nour Eldeen M. Khalifa, Mohamed Hamed N. Taha, Dalia Ezzat Ali, Adam Slowik (Senior Member, IEEE), AND Aboul Ella Hassanien, “Artificial Intelligence Technique for Gene Expression by Tumor RNA-Seq Data: A Novel Optimized Deep Learning Approach”.
  12. Chung-Jui Tu, Li-Yeh Chuang, Jun-Yang Chang, and Cheng-Hong Yang, Member, IAENG,” Feature Selection using PSO-SVM”.
  13. Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, Hayit Greenspan, “GAN-based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion”.
  14. Adrian Yijie Xu, Building a Malaria Classifier with Keras: Background Implementation, Medium: Gradient Crescent. <https://medium.com/gradientcrescent/building-a-malaria-classifier-withkeras-background-implementation-d55c32773afa>.