[1]V. Kalaimani, [2]Dr.R. Umagandhi,

Research Article

Ensemble Mutation Weight Convolutional Neural Network (Emwcnn) Classifier For Gene Expression Microarray Data

[1]V. Kalaimani,
Assistant Professor, Department of Computer Science (PG),
PSGR Krishnammal College for Women, Coimbatore.
mail id: kalaimani_95@yahoo.co.in

[2]Dr.R. Umagandhi,
Associate Professor and Head, Department of Computer Technology,
Kongunadu Arts and Science and College, Coimbatore.
mail id: umakongunadu@gmail.com

ABSTRACT: Gene expression data can reflect gene activities and physiological status in a biological system at the transcriptome level. Gene expression data typically includes small samples but with high dimensions and noise. Hybrid Ensemble Feature Selection (HEFS) system is introduced recently for solving feature selection issue. In the recent work, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Recursive Neural Networks (RNN) classifiers have been introduced for classification. These Machine Learning (ML) techniques are results in low accuracy and performance metrics because of single classifier. Thus Deep Learning (DL) architecture, achieved better results than using ML algorithms in terms of testing accuracy and performance metrics. In this paper, Ensemble Mutation Weight Convolutional Neural Network (EMWCNN) classifier is introduced for classification of gene expression data. To address the performance degradation of single ML classifiers, the construction of an EMWCNN classifier for multi-platform fusion is proposed for gene expression data. In the training stage, multiple CNNs are trained as weak learners and fine-tuned to minimize the weighted error. Meanwhile, optimum weights are selected via adaptive mutation operator. Subsequently, combine the predictions of multiple CNNs to achieve a boosting-based prediction in the reference stage. EMWCNN method proves to be very efficient in classification of gene expression from microarray data, as it involves the process of considering opinion from multiple base classifiers, as opposed to the single classifier method. Experimentation is carryout on four Gene Expression Microarray (GEM) datasets (Prostate cancer, Small Round Blue Cell Tumors (SRBCT), Leukemia, and Lymphoma). Experimental results verify that the proposed EMWCNN method shows improved results with respect to precision, recall, accuracy and Area UnderCurve (AUC) when compared to conventional classifiers.

INDEX TERMS: gene expression, microarray data, Feature selection, classification, Deep Learning (DL), GEM data, and Ensemble Mutation Weight Convolutional Neural Network (EMWCNN).

1. INTRODUCTION

Microarrays are a well established technology to analyze the expression of many genes in a single reaction whose applications range from cancer diagnosis to drug response. They are matrices, where

known samples of DNA, cDNA, or oligonucleotides, called probes, combine with mRNA sequences. The expression level of genes is given by the amount of mRNA bounding to each entry. The aim is to find either sets of genes that characterize particular disease states or experimental condition or highly correlated genes that share common biological features. Microarray numerical data coming out from experiments are normalized and analyzed [1].

In microarray data, the ratio of number of genes (features) to the number of patients (samples) is much skewed which results in the well-known curse-of-dimensionality problem [1]. This further imposes two self-inflicting limitations on any proposed model: (i) processing all the data is not always feasible; and (ii) processing only a subset of data may result in loss of information, overfitting, and local maxima. These two limitations directly impact the accuracy and reliability of any machine learning model. To address the curse-of-dimensionality, a lot of research has been done in the past to identify the most impactful feature subset [2–4]. Both evolutionary as well as statistical methods have been proposed in the literature for this purpose.

Feature Subset Selection (FSS) techniques like Minimum Redundancy Maximum Relevance (mRMR), Joint Mutual Information (JMI), and Joint Mutual Information Maximization (JMIM) are amongst the most prominent statistical methods [5] while advanced approaches like Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Deep Neural Networks (DNN), Transfer Learning, mining techniques, etc. have also been shown in the literature to produce highly accurate results [6–8]. The microarray data classification process is typically carried out in two major phases: (i) Feature Selection: this phase focuses on selecting the most relevant features from otherwise a huge dataset to reduce noise, computational overheads, and overfitting. (ii) Classifier Training: this phase builds a model from the selected features to classify a given microarray sample accurately and reliably.

Classifiers are the core component of microarray data analysis. In the literature, statistical approaches like weighted voting scheme, nearest neighbor classification, discrimination methods and least square and logistic regression were used to develop the classifier model for gene expression data. These statistical approaches usually result in an inflexible classification system that is unable to classify a sample, if the expressions of genes are slightly different from the predefined profile. There is also a variety of machine learning-based classification methods that can be used with genes feature selection for improving classification accuracy results. However, there is still room for improving the accuracy of cancer classification from microarray data.

There are some other works that have been employed the microarray data for cancer detection and diagnosis. The recent work introduced a Deep Learning (DL) method such as Deep Neural Network (DNN), Convolutional Neural Network (CNN) that aims to increase accuracy by using microarray data for cancer detection. Microarray datasets with different DL methods for diagnosing cancer, however ensemble DL consistently performs well in classifying biological data than the single DL methods. The motivation for this study is to utilize robust ensemble methods that are less sensitive to the selection of genes and are capable of removing the uncertainties of gene expression data.

Ensemble methodology is an efficient technique that has increasingly been adopted to combine multiple learning algorithms to improve overall prediction accuracy [9]. These ensemble techniques have the advantage to alleviate the small sample size problem by averaging and incorporating over multiple classification models to reduce the potential for overfitting the training data. In this way the training data set may be used in a more efficient way, which is critical to many bioinformatics applications with small

¹V. Kalaimani, ²Dr.R. Umagandhi,

sample size. Much research has shown the promise of ensemble learning for improving the accuracy in classifying data under uncertainties. However, a necessary and sufficient condition for an ensemble to outperform its individual members is that the base classifiers should be accurate and diverse.

In this paper, Ensemble Mutation Weight Convolutional Neural Network (EMWCNN) classifier is introduced for classification of gene expression data. Ensemble of classifiers is a set of base CNN classifiers that classify a new gene expression based on the weights. Homogenous CNN classifiers are combined for accurate prediction of the presence or absence of disease. The major intent of this EMWCNN classifier is to perform ensemble classifier which are well associated with the label and separate from every other. It gives improved results for cancer dataset when compared to traditional classifiers. Experimental results verify that the EMWCNN classifier shows improved results regarding precision, recall, accuracy and Area Under Curve (AUC) when compared to conventional classifiers. The remaining part of article is emphasized as follows: Section 2 surveys the previous classification and deep learning methods for gene expression data. Section 3 describes the EMWCNN methodology for choosing and classifying the gene samples. Section 4 illustrates the efficiency of traditional classifiers and proposed EMWCNN classifier with benchmark datasets. Finally, Section 5 concludes the entire discussion and suggests an extension of this work.

2. LITERATURE REVIEW

Vanitha et al [10] presented an effective method for gene classification using Support Vector Machine (SVM). SVM is a supervised learning algorithm capable of solving complex classification problems. Mutual information (MI) between the genes and the class label is used for identifying the informative genes. The selected genes are utilized for training the SVM classifier and the testing ability is evaluated using Leave-one-Out Cross Validation (LOOCV) method. The performance of the proposed approach is evaluated using two cancer microarray datasets. From the simulation study it is observed that the proposed approach reduces the dimension of the input features by identifying the most informative gene subset and improve classification accuracy when compared to other approaches.

Sreepada et al [11] proposed a microarray classification is done in two phases. In the first phase, a hybrid approach of Genetic Algorithm (GA) and Principal Component Analysis (PCA) is used for extracting relevant features. In the second phase, Probabilistic Neural Network (PNN) is used as the classifier and GA is implemented to optimize the topology of the PNN. The datasets used in the experiment are Colon Tumor, Diffuse Large B-Cell Lymphoma (DLBCL) and Leukaemia (ALL and AML). The proposed technique gave efficient results for the datasets used.

Li et al [12] introduced a more efficient implementation of Linear Support Vector Machines(LSVMs) and improve the Recursive Feature Elimination(RFE) strategy and then combine them together to select informative genes. Besides, a simple resampling method is introduced to preprocess the datasets, which makes the information distribution of different kinds of samples balanced and the classification results more credible. Extensive experiments are conducted on six most frequently used microarray datasets in this field, and the results show that the proposed methods have not only reduced the time consumption greatly but also obtained comparable classification performance.

Potharaju and Sreedevi [13] proposed a Distributed Feature Selection (DFS) strategy using Symmetrical Uncertainty(SU) and Multi Layer Perceptron(MLP) by distributing across the multiple clusters. Each

cluster is equipped with finite number of features in it. MLP is employed over each cluster and based on the highest accuracy and lowest Root Mean Square Error Rate (RMS) dominant cluster is nominated.

Zahoor and Zafar [14] presented a warzone inspired "Infiltration Tactics" based Optimization algorithm (ITO)—not to be confused with the ITO algorithm based on the Itõ Process in the field of Stochastic calculus. The proposed ITO algorithm combines parameter-free and parameter-based classifiers to produce a High-Accuracy-High-Reliability (HAHR) binary classifier. The algorithm produces results in two phases: (i) Lightweight Infantry Group (LIG) converges quickly to find non-local maxima and produces comparable results (i.e., 70 to 88% accuracy) (ii) Followup Team (FT) uses advanced tuning to enhance the baseline performance (i.e., 75 to 99%). Every soldier of the ITO army is a base model with its own independently chosen Subset selection method, pre-processing, and validation methods and classifier. The successful soldiers are combined through heterogeneous ensembles for optimal results.

Mallick et al [15] presented method of classification to understand the convergence of training Deep Neural Network (DNN). The assumptions are taken as the inputs do not degenerate and the network is over-parameterized. Also the number of hidden neurons is sufficiently large. Authors in this piece of work have used DNN for classifying the gene expressions data. The dataset used in the work contains the bone marrow expressions of 72 leukemia patients. A five-layer DNN classifier is designed for classifying Acute Lymphocyte (ALL) and Acute Myelocytic (AML) samples. The network is trained with 80% data and rest 20% data is considered for validation purpose. The different types of computer-aided analyses of genes can be helpful to genetic and virology researchers as well in future generation.

Liao et al [16] proposed a novel Multi-Task Deep Learning (MTDL) method to solve the data insufficiency problem. Since MTDL leverages the knowledge among the expression data of multiple cancers to learn a more stable representation for rare cancers, it can boost cancer diagnosis performance even if their expression data are inadequate. The experimental results show that MTDL significantly improves the performance of diagnosing every type of cancer when it learns from the aggregation of the expression data of twelve types of cancers.

Kong and Yu [17] proposed a Forest Deep Neural Network (fDNN) classifier to integrate the deep neural network architecture with a supervised forest feature detector. Using this built-in feature detector, the method is able to learn sparse feature representations and feed the representations into a neural network to mitigate the overfitting problem. Simulation experiments and real data analyses using two RNA-seq expression datasets are conducted to evaluate fDNN's capability. The method is demonstrated a useful addition to current predictive models with better classification performance and more meaningful selected features compared to ordinary Random Forests(RFs) and DNN.

Elbashir et al [18] proposed a lightweight Convolutional Neural Networks (CNN) architecture for breast cancer classification using gene expression data downloaded from Pan-Cancer Atlas using "Illumina HiSeq" platform. The downloaded gene expression data is preprocessed and then transformed into 2D-images. Started the preprocessing by removing the outlier samples, which are determined based on the Array-Array Intensity Correlation (AAIC), which defines a symmetric square matrix of Spearman correlation. Then a normalization process is applied on the gene expression data to ensure that we can infer the expression level from it correctly and avoid biases in the expression measures. Finally, filtering is applied on the data. Model selection or a parameters search strategy is conducted to choose the values of the CNN hyper-parameters that give optimal performance. Experiments show that proposed method achieves an accuracy of 98.76%, which is the highest compared to other competing methods.

[1]V. Kalaimani, [2]Dr.R. Umagandhi,

Sevakul et al [19] presented a transfer learning procedure for cancer classification, which uses feature selection and normalization techniques in conjunction with s sparse autoencoders on gene expression data. While classifying any two tumor types, data of other tumor types were used in unsupervised manner to improve the feature representation. The performance of algorithm was tested on 36 two-class benchmark datasets from the Gene Expression Machine Learning Repository (GEMLeR) repository. On performing statistical tests, it is clearly ascertained that algorithm statistically outperforms several generally used cancer classification approaches. The deep learning based molecular disease classification can be used to guide decisions made on the diagnosis and treatment of diseases, and therefore may have important applications in precision medicine.

Xia et al [20] introduced a Convolutional Neural Network(CNN) based multi-model ensemble method for cancer prediction using DNA methylation data. Five basic machine learning methods are choose as the first stage classifiers and conduct prediction individually. Then, a CNN is used to find the high-level features among the classifiers and gives a credible prediction result. Experimental results on three DNA methylation datasets of Lung Adenocarcinoma, Liver Hepatocellular Carcinoma and Kidney Clear Cell Carcinoma show the proposed ensemble method can uncover the intricate relationship among the classifiers automatically and achieve better performances.

3. PROPOSED METHODOLOGY

In cancer classification, single Machine Learning (ML) techniques seems to be not capable of ensuring optimal results in terms of both predictive performance and stability, thus ensemble approaches has been focused by researchers by the combination of different multiple CNNs. Firstly, Feature subset produced by Hybrid Ensemble Feature Selection (HEFS) algorithm is used for classification. HEFS algorithm combines the three feature selection algorithms such as filter, embedded method and wrapper. Then, the classification is performed based on selected features from HEFS algorithm. The classification is performed based on the Ensemble Mutation Weight Convolutional Neural Network (EMWCNN) classifier for gene expression data. Validate its superiority with four Gene Expression Microarray (GEM) datasets. Experimental results verify that the proposed classifier shows improved results with respect to precision, recall, accuracy and Area Under Curve (AUC). The overall framework of the proposed HEFS feature selection algorithm is shown in the figure 1.

3.1. HYBRID ENSEMBLE FEATURE SELECTION (HEFS) ALGORITHM

Hybrid Ensemble Feature Selection (HEFS) algorithm [24] is introduced which combines the three feature selection algorithms such as filter by Score-Based Criteria Fusion (SCF) and embedded method by Fuzzy Elephant Herding Optimization (FEHO), and Support Vector Machine- t (SVM-t). The results of these methods are aggregated via the Weighted Majority Voting (WMV) [24].

3.1.1. FILTER FEATURE SELECTION

Filter feature selection methods use statistical techniques to evaluate the relationship between each input variable and the target variable, and these scores are used as the basis to choose (filter) those input variables that will be used in the model. In Score-Based Criteria Fusion (SCF), the relevance estimation measure consists of two parts, i.e., Symmetrical Uncertainty (SU) and ReliefF, combined by using a specific fusion method. According to the combined objects, the fusion methods are divided into two categories: score-based multicriterion fusion and ranking-based multicriterion fusion [21-22]. Finally features are sorted according to their values in the final score vector. Concretely, combination algorithm

fuses two score vectors from two basis criteria by multiplying the weight parameter. In addition, the search strategy of SCF algorithm follows the incremental forward selection.

3.1.2. WRAPPER FEATURE SELECTION

Wrapper feature selection by evaluating a subset of features using a machine learning algorithm that employs a search strategy to look through the space of possible feature subsets, evaluating each subset based on the quality of the performance of a given algorithm. It follows a Fuzzy Elephant Herding Optimization (FEHO) by evaluating all the possible combinations of features against the evaluation. FEHO method is inspired by the herding behavior of elephant group [23]. As this work is focused on improving the FEHO updating process, in the following subsection, provide further details of the FEHO updating operator as it was originally presented. For details regarding the FEHO separating operator for optimal selection of features, see the literature [23].

3.1.3. EMBEDDED FEATURE SELECTION

Embedded methods complete the feature selection process within the construction of the machine learning algorithm itself. It is performed based on the Support Vector Machine- t (SVM-t) statistics to choose analytical features from the dataset. SVM make use of only the information of support vectors in the direction of creating the maximal partition hyper plane and find the classes for every dataset. The general two-sample t-statistic is utilized in the direction of evaluates the important variation among two classes. It is perceptive to choose features by way of the improved accuracy as the feature set [24].
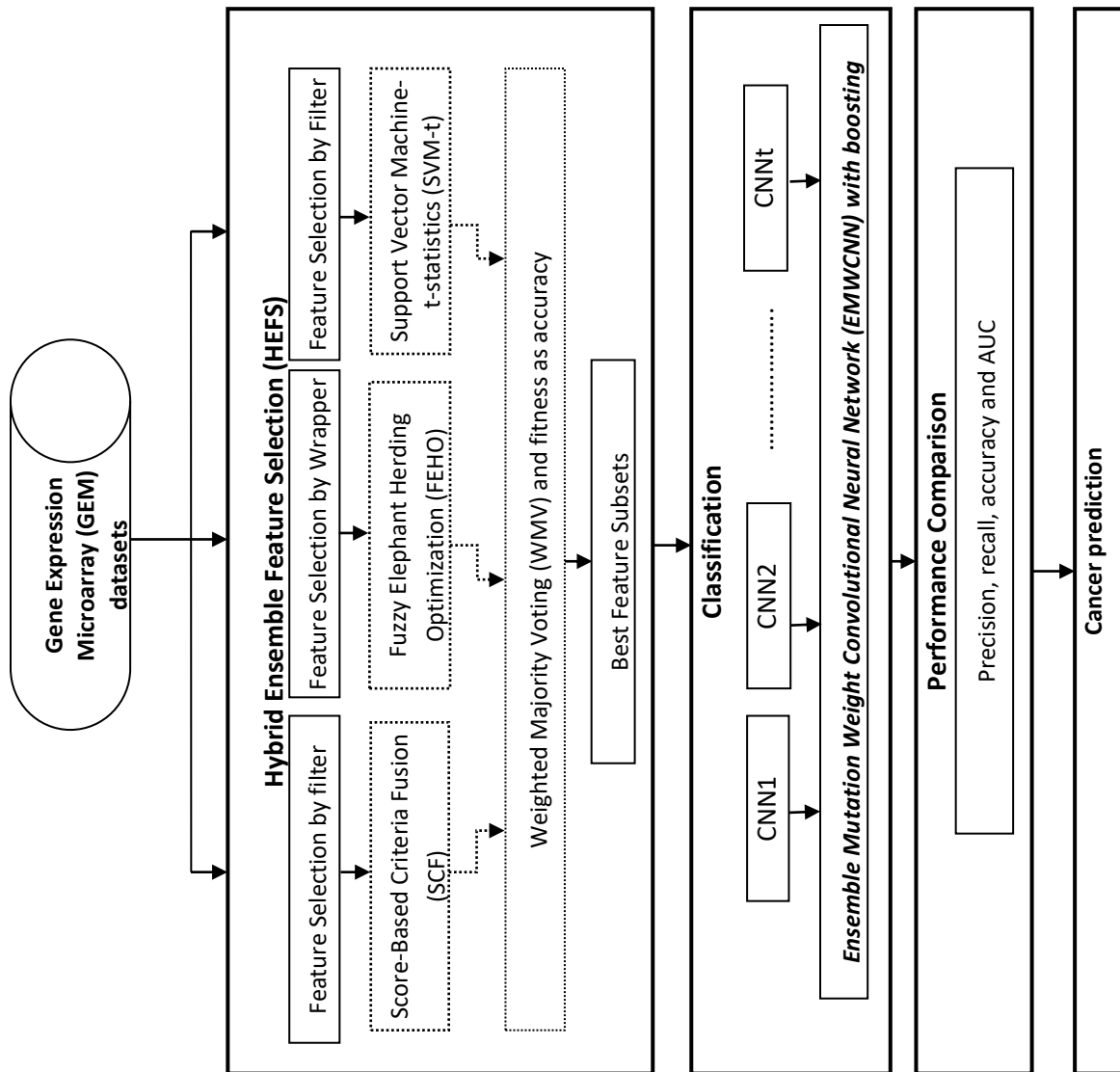
[1]V. Kalaimani, [2]Dr.R. Umagandhi,

FIGURE 1. OVERALL RESEARCH FRAMEWORK OF THE PROPOSED OSCF BASED FEATURE SELECTION

3.1.4. Weighted Majority Voting (WMV)

Prediction accuracy can reinforce the decision of those qualified features, which makes it possible to give more importance to their decision in the vote and consequently may further improve the overall performance than that can be obtained by SMV (where all feature selection methods have identical weights). In WMV, each vote is weighted by the prediction accuracy value of the features via classifier [24].

3.2. FINTESS COMPUTATION

Fitness is computed by combining the classification accuracy and SCF. If both are higher than the features are selected from the GEM dataset. Fitness is calculated from the Classifiers such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Recursive Neural Networks (RNN) and EMWCNN. Selected features are given to classifier, and then accuracy will also be considered with fitness along with

SCF criteria. Classification accuracy is defined as a ratio between the numbers of correctly assigned class labels and the total number of objects to be classified.

3.3. Ensemble Mutation Weight Convolutional Neural Network (EMWCNN) classifier

Ensemble Mutation Weight Convolutional Neural Network (EMWCNN) module is pre-trained and fine-tuned with training samples to learn network parameters. The ensemble weights are determined by multiple EMWCNN having different properties, which are used to collaboratively infer the labels of testing gene expression data in the fusion layer. As illustrated in Figure 1, the proposed architecture comprises three parts, namely input datasets, multi-branch EMWCNN modules, and decision-level fusion classification. All the processed gene expression are stacked into a training set $S \in \mathbb{R}^{N_1 \times N_2 \times Q}$, where Q is the gene expression data sample number, $N_1 \& N_2$ denotes the rows and columns of the sample. The features selected data are fed into each EMWCNN to train the hierarchical structure in an unsupervised manner. Then, weights are generated according to the properties of the EMWCNNs with ensemble. For the proposed ensemble EMWCNNs, they are operated in a supervised manner as the training labels are indispensable.
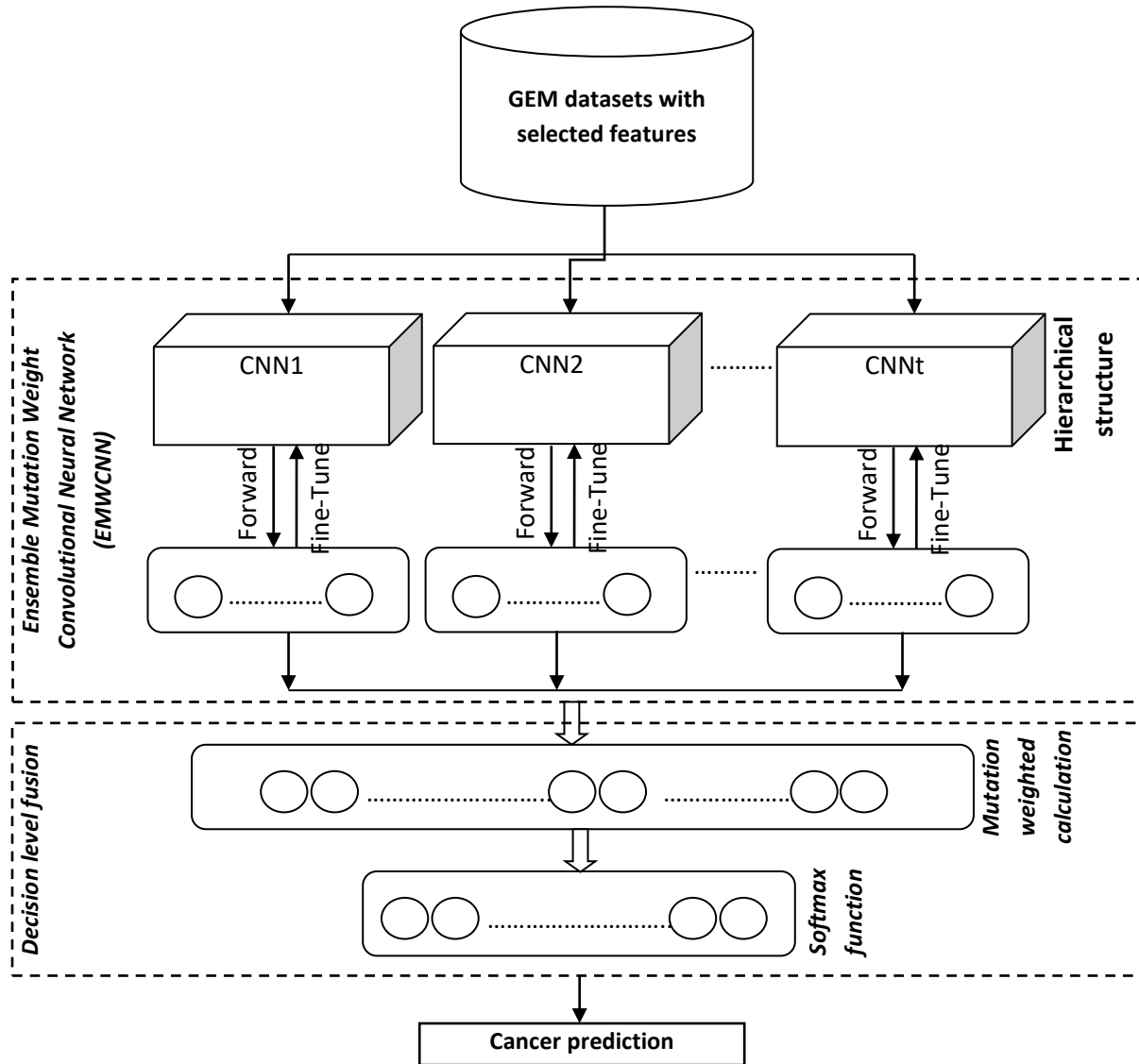
[1]V. Kalaimani, [2]Dr.R. Umagandhi,

FIGURE 2. ARCHITECTURE OF EMWCNN FOR CLASSIFICATION

As shown in Figure 2, T EMWCNNs are of the same structure, but with different network parameters. Generally, EMWCNNs comprises a convolutional layer, an activation function, a pooling layer, and a fully-connected layer. To elaborate, the preprocessed input $S(t) \in \mathbb{R}^{N_1 \times N_2 \times Q}$, is convoluted with randomly initialized kernels F and a bias term b is added to the resultant feature maps. Instead of a sigmoid function in general, the ReLU activation function $g(x) = max(0, x)$ is a more advisable alternative to boost the nonlinear pointwise network. It has been proven to speed up the convergence significantly and increase the descriptive ability of the network [25]. Finally, the max-pooling layer is utilized to reduce the sensitivity to small input shifts. Then, given T CNNs, the kth layer output state of the tth branch can be formulated as follows by equation (1) [26],

$$L_k^t = pool\left(ReLU\left(L_{k-1}^t \otimes F_k^t + b_k^t\right)\right) \text{ (1)}$$

where $\otimes$ is the 2D convolutional operation, and pool($\cdot$) denotes the pooling operation. The network parameters, that is, $\theta = \{F_k^t, b_k^t, t = 1, \dots, T, k = 1, \dots, K\}$ are iteratively optimized by minimizing the classification error over the training gene expression data. Commonly, the Stochastic Gradient Descent (SGD) algorithm is adopted to achieve this. In SGD, the network parameters are learned by the derivatives of F and b [26]. The learning performance is related to the slight learning rate ω during optimization. Once the single EMWCNNs parameters are well trained and the multiple EMWCNNs are weighted with ensemble, the decision-level fusion recognition can be executed for given multiple input gene vectors $\{x_t, t = 1, \dots, T\}$, which yields [26],

$$H(Y_{fusion}) = \arg\max_{i \in \{1,\dots,C\}} \left[\sum_{t=1}^{T} \beta_t \Pr(Y_t = i|\theta_t)\right] \quad (2)$$

$$= \arg\max_{i \in \{1,\dots,C\}} \left[\sum_{t=1}^{T} \frac{\exp(MW^t x_t + b^t)}{\sum_{i=1}^{C} \exp(MW_i^t x_t + b_i^t)}\right] \quad (3)$$

where $\{\beta_t\}$ are produced via ensemble, $\theta_t = \{W_t, b_t, t = 1, \dots T\}$ is the network parameter set of the decision- making logistic regression layer, MW_t and b_t are the mutation weights and biases, and C is the total number. MW_t is generated via the Diversity based Mutation operator [27].

In this Diversity based Mutation scheme, more probability is used for mutation to a weight that has less population-wise diversity. To implement, first the variance of weight values of each weight across the population members is computed and weights are sorted in ascending order of variance. Thereafter, an exponential probability distribution $(p(i) = \lambda \exp(-\lambda i)$ for $i \in [0, n-1]$ is used. To make the above a probability distribution, $\bar{\lambda}$ is used by finding the root of the following equations (4-5) for a fixed n (weight values) [27],

$$\lambda \exp(-n\lambda) - \exp(-\lambda) - \lambda + 1 = 0 \quad (4)$$

$$MW_t = l = \frac{1}{\lambda}\log(1 - u(1 - \exp(-n\bar{\lambda})))) \quad (5)$$

The label information can be therefore deduced by the mutation weighted probability of each class. The processing branches are composed of T CNNs, but the intercepted input samples and EMWCNNs structures are not identical. The difference between them is measured by mutation weights with boosting in this subsection. AdaBoost is introduced to control the selection of multiple weak classifiers to form a strong to produce a more convincing recognition. As the resource data are obtained from the same category of samples, boosting is performed to train multiple EMWCNNs and assign weights in the reference stage. However, the key to the algorithm scheme is to minimize the training error with a pre-assigned number of iterations. In the proposed EMWCNNs Boosting, training stage multiple EMWCNNs are trained as weak learners and fine-tuned to minimize the weighted error. Meanwhile, optimum weights are selected. Subsequently, combine the predictions of multiple EMWCNNs to achieve a boosting-based prediction in the reference stage. The proposed EMWCNN Boosting is described in detail in Algorithm 1. Notably, the final predictive inference is determined via a majority voting strategy.

Algorithm 1: EMWCNN Boosting Algorithm

Input: Labelled Training Samples $S \in \mathbb{R}^{N_1 \times N_2 \times Q}$, Where Q Is The Gene Expression Data Sample Number, N_1 & N_2 Denotes The Rows And Columns Of The Sample, Labels Y_{tr}, Iteration Number T,

[1]V. Kalaimani, [2]Dr.R. Umagandhi,

MWCNN Optimization Parameters ω, Label Number C

Training: Initialize the distribution uniformly : $D_t^{(q)} = 1/Q$

Loop: For $t = 1, \ldots.T$

Step 1: Train CNN t with S and Y_{tr} while randomly selecting ω_t

Step 2: Predict the labels of validation set by the maximum probability of output perceptrons $Y_{pre,q} = \arg\max_{i=1,\ldots C}\{\Pr(S_q = i)\}, q = 1, \ldots Q$

Step 3: Calculate the error on $D_t^{(q)}, \varepsilon_t = b_t D_t^{(q)}$, $b_t = \max_{i\neq c}\Pr(Y_{pre,q} = i) - \Pr(Y_{pre,q} = c)$ and c is the actual label

Step 4: Choose $\beta_t \in \mathbb{R}$ as $\beta_t = \frac{1}{2}\log\left[(C-1)\frac{1-\varepsilon_t}{\varepsilon_t}\right]$,

Step 5: Update distribution $D_{t+1}^{(q)} = D_t^{(q)}\exp(\beta_t b_t MW_t)/Z_t$, Z_t is the normalization factor to make $\sum_q^Q D_{t+1}^{(q)} = 1$

End

Output: $\{\beta_t\}_{t=1}^T$,Then the classification is subsequently executed with mutation weighted probability and majority voting

4. RESULTS AND DISCUSSION

In order to validate the algorithms of Sort Aggregation -EFS, and proposed HEFS algorithm, experiments are conducted on the following four gene expression microarray datasets.

4.1. DATASET DESCRIPTIONS

Experiments are conducted on the following four gene expression microarray datasets:

Prostate cancer

Prostate data consisted of 102 samples where 50 samples are prostate tumors and 52 sample is normal. Each sample contains 10509 genes. This dataset can be downloaded at http://www.gems-system.org/.

SRBCT data

Small Round Blue-Cell Tumor (SRBCT) dataset consists of 83 samples, each containing 2,308 genes. The tumors are Burkett's Lymphoma (BL), the Ewing Family of tumors, NeuroBlastoma (NB) and RhabdoMyo Sarcoma (RMS). There are 63 samples for training and 20 samples for testing. The training set consists of 8, 23, 12 and 20 samples of BL, EWS, NB and RMS respectively. The test set consists of 3, 6, 6 and 5 samples of BL, EWS, NB and RMS respectively. This dataset can be downloaded at http:// www.biolab. si/ supp /bi-cancer/projections/info/ SRBCT.html

Leukemia

Leukemia dataset holds expression levels of 7129 genes in use over 72 models. This dataset is of the similar type as the colon cancer dataset and can 56 consequently be used for the similar kind of experiments. It consists of 72 sample, 25 samples of Acute Myeloid Leukemia (AML) and 47 samples of Acute Lymphoblastic Leukemia (ALL). The source of the gene expression measurements is taken from 63 bone marrow samples and 9 peripheral blood samples. This dataset can be downloaded at http://cilab.ujn.edu.cn/datasets.html.

Lymphoma

Lymphoma has two distinct tumor subtypes of B-DLCL are germinal center B cell-like DLCL and activated B cell-like DLCL. Lymphoma dataset consists of 24 samples of germinal center B-like and 23 samples of activated B-like. Lymphoma dataset holds 42 samples resultant from diffuse large B-cell lymphoma (DLBCL) and 9 samples from Follicular Lymphoma (FL) later than 11 samples from Chronic Lymphocytic Leukemia (CLL) downloaded from http://csse.szu.edu.cn/staff/zhuzx/Datasets.html. The whole dataset hold 4026 genes. This dataset contains some missing values which are imputed using K-Nearest Neighbor method. Table 1 briefly summarizes these datasets.

TABLE 1. GENE DATASETS CHARACTERISTICS

DATASETS	#GENE	#INSTANCE	#CLASS
LEUKEMIA	7129	72	2
LYMPHOMA	4026	62	3
PROSTATE CANCER	10509	102	2
SMALL ROUND BLUE CELL TUMORS (SRBCT)	2308	83	4

4.2.   PERFORMANCE METRICS

To evaluate the performance of proposed feature selection algorithm we use four well known classifiers, namely, Support Vector Machines (SVM) , Recursive Neural Networks (RNN), CNN, and proposed Ensemble Mutation Weight Convolutional Neural Network (EMWCNN). The performance of proposed Hybrid Ensemble Feature Selection (HEFS) algorithm is compared with standard EFS with respect to four benchmark datasets such as Prostate cancer, Small Round Blue Cell Tumors (SRBCT), Leukemia, and Lymphoma via MATLAB environment. These methods are assessed using the classification metrics like precision, recall, accuracy and Area Under Curve (AUC). Four effective measures calculated from confusion matrix output from the Table 2, which are True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN).

Precision is also referred to as positive predictive value as shown in equation (6)

[1]V. Kalaimani, [2]Dr.R. Umagandhi,

$$Precision=TP/(TP+FP) \quad (6)$$

Recall also known as sensitivity is a true positive rate that is the ratio of True Positive (TP) to the sum of True Positives (TP) and False Negatives (FN) as shown in equation (7)

$$Recall=TP/(TP+FN) \quad (7)$$

Accuracy works by considering the number of correctly classified samples to the ratio of the total number of test samples as shown in equation (8)

$$Accuracy=(TP+TN)/(TP+TN+FP+FN) \quad (8)$$

Another measure is the Area Under Curve (AUC) wherein the curve is the receiver operating characteristic (ROC-curve) curve. The ROC-curve is the graphical plot of the True Positive Rate (TPR) versus the False Positive Rate (FPR) for a binary classifier as its discrimination threshold is varied.

TABLE 2.CONFUSION MATRIX

Total population		Predicted class	
		Prediction Positive	Prediction Negative
Actual class	Condition Positive	True Positive (TP)	False Negative (FN)
	Condition Negative	False Positive (FP)	True Negative (TN)

4.3. COMPARISON OF PERFORMANCE METRICSVS.METHODS

To evaluate the performance of proposed feature selection algorithm we use four well known classifiers, namely, Support Vector Machines (SVM) , Recursive Neural Networks (RNN), CNN, and proposed Ensemble Mutation Weight Convolutional Neural Network (EMWCNN). The following figures show the results of various metrics with four benchmark datasets.
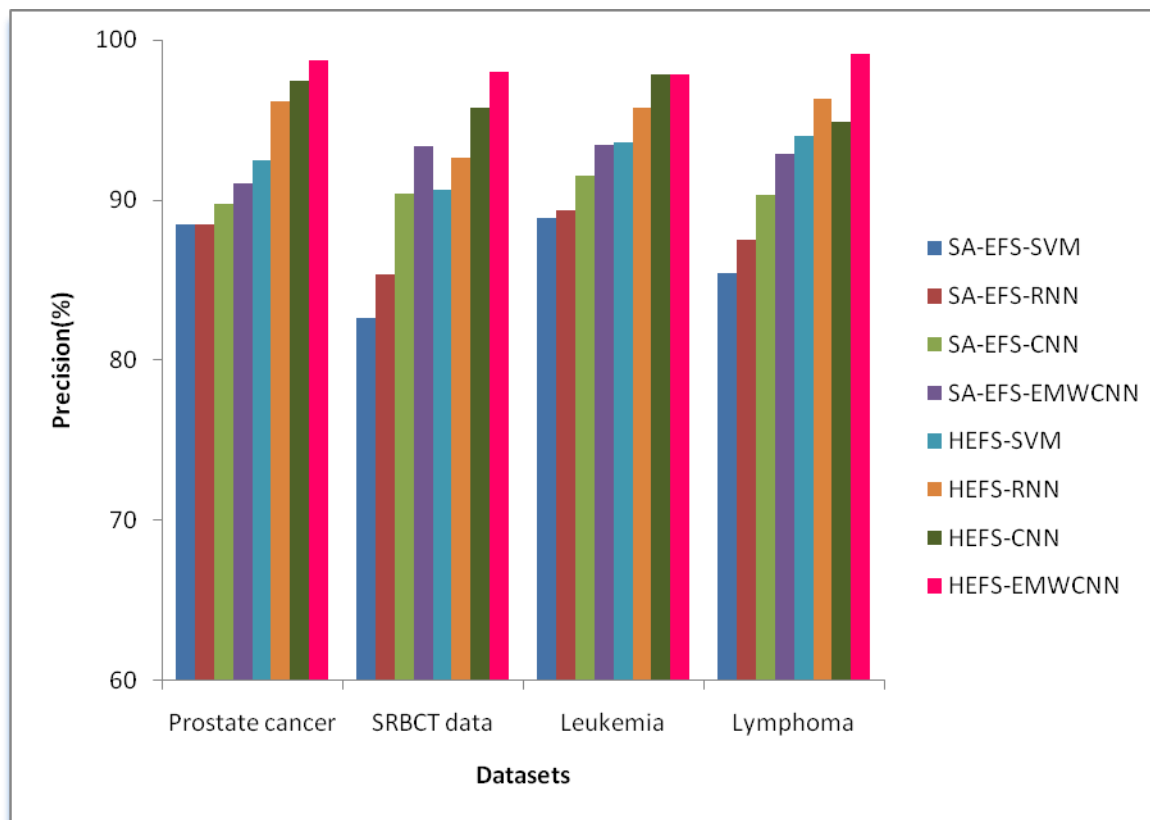
FIGURE.3 PRECISION RESULTS COMPARISON VS. GEM DATASETS

Figure 3 shows the comparison of proposed HEFS feature selection and existing SA-EFS feature selection in terms of precision. These feature selection methods have been experimented using three classifiers such as KNN, SVM, CNN and EMWCNN. In x-axis, four benchmark datasets have been considered and the precision results are shown in the y-axis. From this analysis, it is analyzed that the proposed HEFS feature selection technique can achieve better precision than other existing EFS. From classifiers (SVM, RNN, CNN and EMWCNN), proposed HEFS–EMWCNN gives highest precision of 98.7500% for prostate cancer dataset which is 7.72% higher than SA-EFS-EMWCNN (See Table 3). Similarly, the precision value of HEFS–CNN for prostate cancer dataset is 7.73% higher than SA-EFS-CNN, HEFS-RNN is 7.74% higher than SA-EFS– RNN, and HEFS-SVM is 4.40% higher than SA-EFS–SVM(See Table 3).
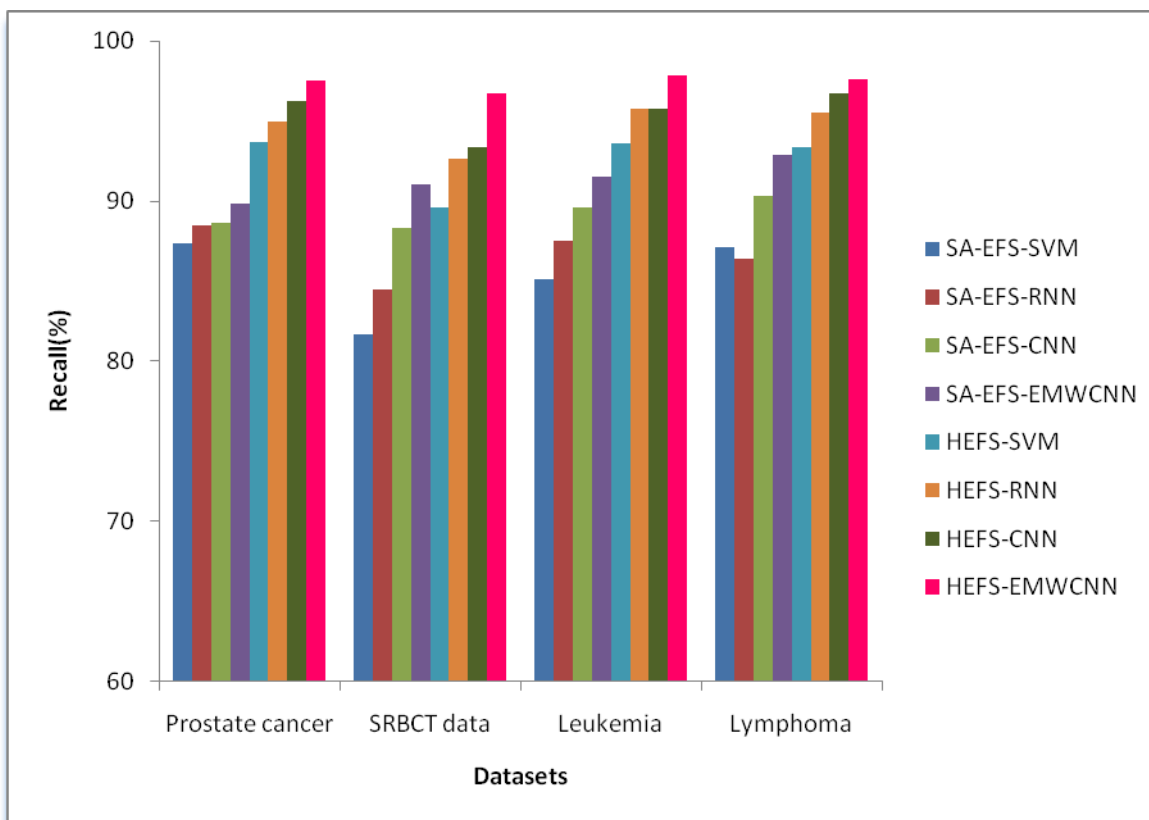
[1]V. Kalaimani, [2]Dr.R. Umagandhi,

FIGURE.4 RECALL RESULTS COMPARISON VS. GEM DATASETS

Figure 4 shows the comparison of proposed HEFS feature selection and existing SA-EFS feature selection in terms of recall. These feature selection methods have been experimented using three classifiers such as KNN, SVM, CNN and EMWCNN. In x-axis, four benchmark datasets have been considered and the recall results are shown in the y-axis. From this analysis, it is analyzed that the proposed HEFS feature selection technique can achieve better recall than other existing EFS. From classifiers (SVM, RNN, CNN and EMWCNN), proposed HEFS–EMWCNN gives highest recall of 97.5310% for prostate cancer dataset which is 7.658% higher than SA-EFS-EMWCNN (See Table 3). Similarly, the recall value of HEFS–CNN for prostate cancer dataset is 7.642% higher than SA-EFS-CNN, HEFS-RNN is 6.538% higher than SA-EFS–RNN, and HEFS-SVM is 6.329% higher than SA-EFS–SVM (See Table 3).
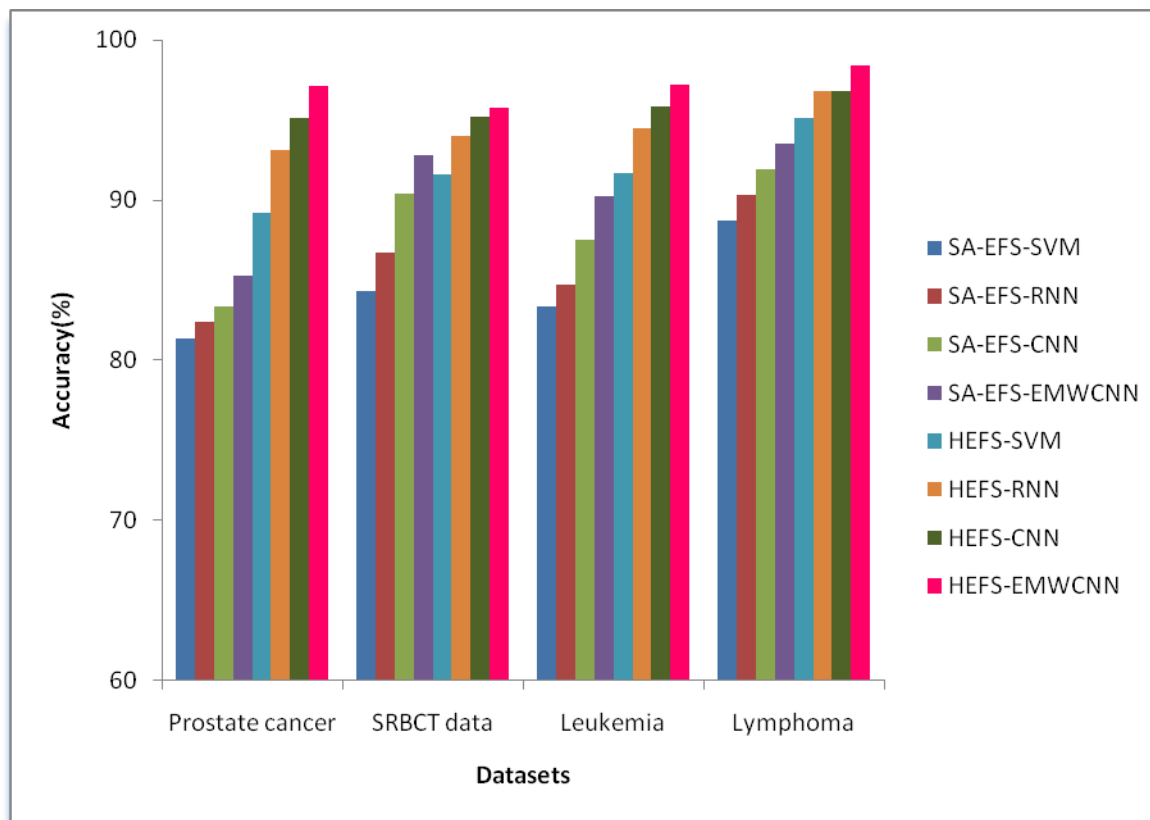
FIGURE.5 ACCURACY RESULTS COMPARISON VS. GEM DATASETS

Figure 5 shows the comparison of proposed HEFS feature selection and existing SA-EFS feature selection in terms of accuracy. These feature selection methods have been experimented using three classifiers such as KNN, SVM, CNN and EMWCNN. In x-axis, four benchmark datasets have been considered and the accuracy results are shown in the y-axis. From classifiers (SVM, RNN, CNN and EMWCNN), proposed HEFS–EMWCNN gives highest accuracy of 97.0873 % for prostate cancer dataset which is 11.7932% higher than SA-EFS-EMWCNN (See Table 3). Similarly, the accuracy value of HEFS–CNN for prostate cancer dataset is 11.7647% higher than SA-EFS-CNN, HEFS-RNN is 10.7843% higher than SA-EFS– RNN, and HEFS-SVM is 7.8431% higher than SA-EFS–SVM (See Table 3).

TABLE 3.RESULTS COMPARISON OF GEM DATASETS VS. CLASSIFIERS

Methods/Metrics	Leukemia Dataset			Prostate Cancer Dataset		
	Precision(%)	Recall(%)	Accuracy(%)	Precision(%)	Recall(%)	Accuracy(%)
SA-EFS-SVM	88.8888	85.1063	83.3333	88.4600	87.3420	81.3725
SA-EFS-RNN	89.3617	87.5000	84.7222	88.4600	88.4620	82.3529
SA-EFS-CNN	91.4894	89.5833	87.5000	89.7400	88.6080	83.3333

[1]V. Kalaimani, [2]Dr.R. Umagandhi,

| SA-EFS-EMWCNN	93.4782	91.4894	90.2778	91.0300	89.8730	85.2941
HEFS-SVM	93.6170	93.6170	91.6666	92.5000	93.6710	89.2156
HEFS-RNN	95.7446	95.7446	94.4444	96.2000	95.0000	93.1372
HEFS-CNN	97.8261	95.7446	95.8333	97.4700	96.2500	95.0980
HEFS-EMWCNN	97.8723	97.8723	97.2222	98.7500	97.5310	97.0873
Methods/Metrics	Lymphoma Dataset			SRBCT Dataset		
	Precision(%)	Recall(%)	Accuracy(%)	Precision(%)	Recall(%)	Accuracy(%)
SA-EFS-SVM	85.4494	87.1240	88.7100	82.6540	81.7085	84.3370
SA-EFS-RNN	87.5144	86.3746	90.3230	85.3472	84.4407	86.7470
SA-EFS-CNN	90.2923	90.2923	91.9350	90.3755	88.3243	90.3610
SA-EFS-EMWCNN	92.8567	92.8567	93.5480	93.3240	91.0183	92.7710
HEFS-SVM	93.9707	93.3334	95.1610	90.6170	89.6230	91.5660
HEFS-RNN	96.3214	95.5557	96.7740	92.6297	92.6345	93.9760
HEFS-CNN	94.8717	96.7180	96.7740	95.7500	93.3928	95.1810
HEFS-EMWCNN	99.0990	97.6190	98.3870	98.0000	96.7263	95.7900
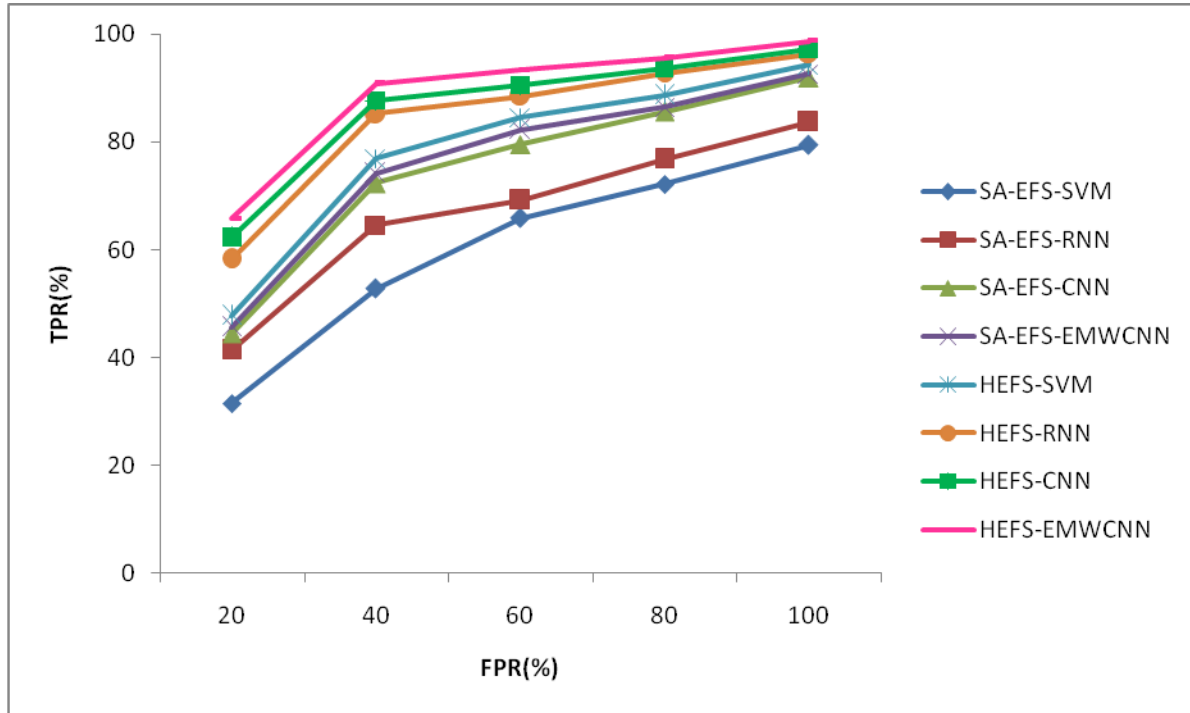
FIGURE.6 (a) ROC CURVE VS. CLASSIFIERS (PROSTATE CANCER DATASET)
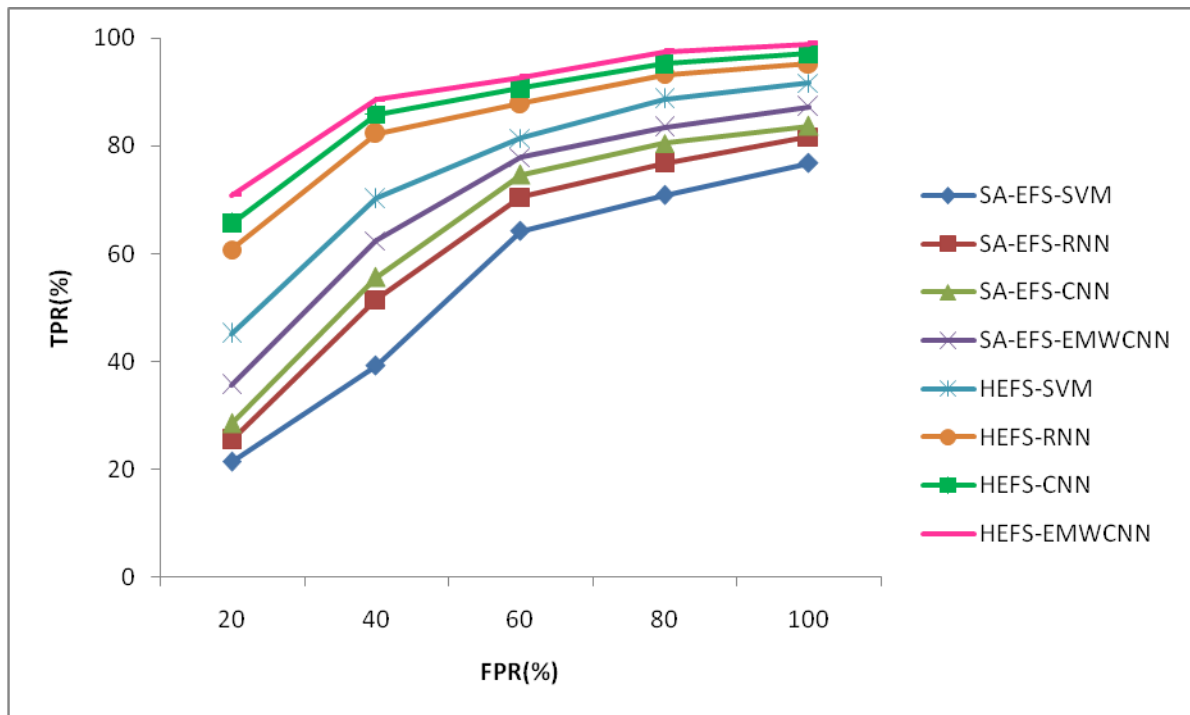
FIGURE.6 (b) ROC CURVE VS. CLASSIFIERS (SRBCT DATASET)
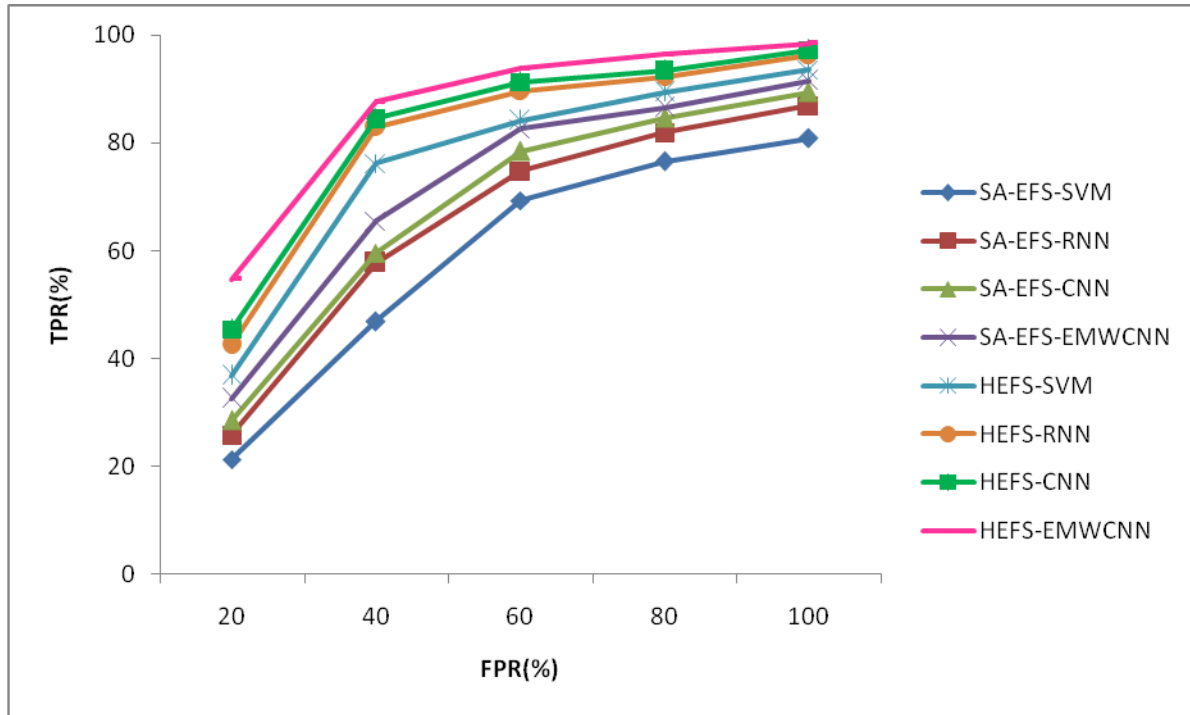
[1]V. Kalaimani, [2]Dr.R. Umagandhi,

FIGURE.6 (c) COMPARISON OF ROC CURVE VS. CLASSIFIERS (LEUKEMIA DATASET)
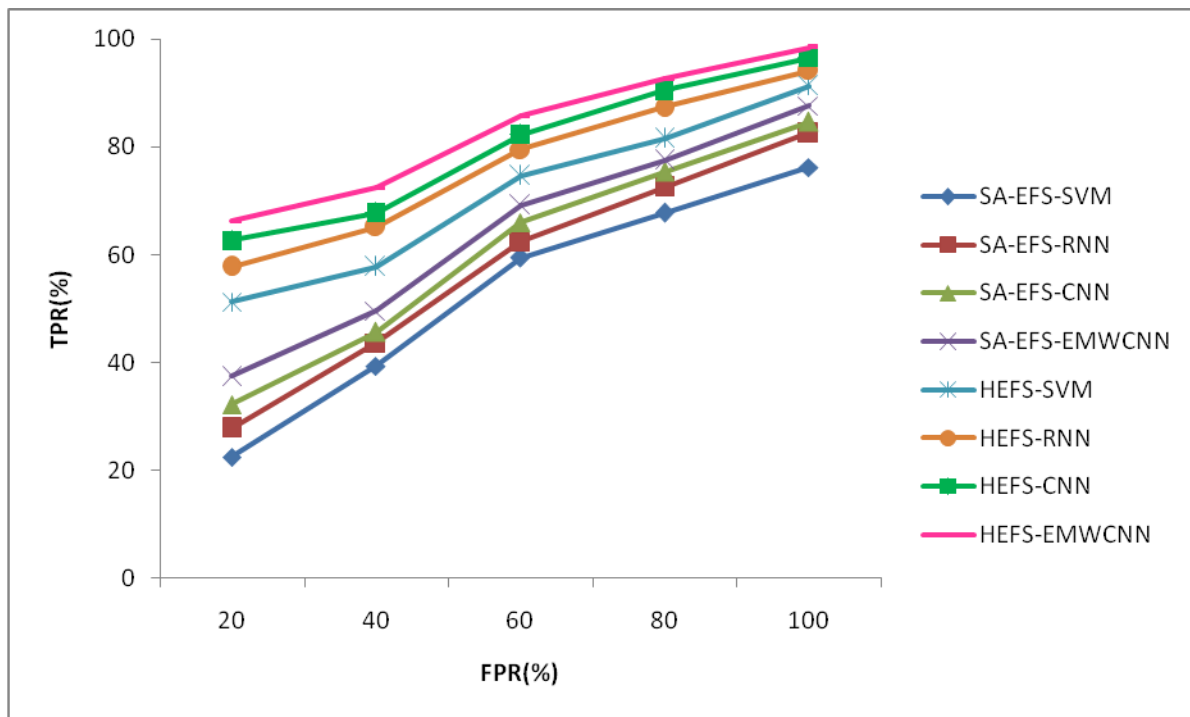

FIGURE.6 (d) COMPARISON OF ROC CURVE VS. CLASSIFIERS (LYMPHOMA DATASET)

In Figure 6(a-d), see that the area under the ROC curves (AUC) obtained by the HEFS–EMWCNN algorithm is the largest when compared to other methods. AUC obtained by the HEFS–EMWCNN algorithm is much higher for prostate cancer dataset when compared to other datasets. In Figure 6(a-d),

see that the effect of HEFS with four classifiers is higher when compared to the effect of SA-EFS with classifiers.

5. CONCLUSION AND FUTURE WORK

Classification approaches have been developed, adopted, and applied to distinguish disease classes at the molecular level using microarray data. Firstly, Feature subset produced by Hybrid Ensemble Feature Selection (HEFS) algorithm is used for classification. Secondly HEFS algorithm combines the three feature selection algorithms such as filter, embedded method and wrapper. Then Ensemble Mutation Weight Convolutional Neural Network (EMWCNN) is introduced for binary class classification and multiclass classification using microarray gene expression data. The most unique advantage of the EMWCNN approach is its ability to explore both the binary and the multiclass underlying relationships between the target features of a given disease classification problem and the involved explanatory gene expression data. EMWCNN classifier is pre-trained and fine-tuned with training samples to learn network parameters. The ensemble weights are determined via adaptive mutation operator in multiple EMWCNN which are used to collaboratively infer the labels of testing gene expression data in the fusion layer. Multiple CNNs are trained as weak learners and fine-tuned to minimize the weighted error. Then, weights are generated according to the properties of the EMWCNNs with ensemble. Subsequently, combine the predictions of multiple EMWCNNs to achieve a boosting-based prediction in the reference stage. This classifier is subsequently executed with mutation weighted probability and majority voting. The experimental results carried out by using gene expression data which are available on the UCI repository. It has been tested with two class and multi-class datasets and compared with the classical classification algorithms such as RNN, SVM, and CNN. The performance of four classifiers are examined using gene expression data and evaluated using Receiver Operating Characteristic (ROC) curve from Prostate cancer, Small Round Blue-Cell Tumor (SRBCT) data, Leukemia and Lymphoma. The result of this work is used to the drug designer for the pathway analysis and disease treatment decisions.

REFERENCES

1. Önskog, J., Freyhult, E., Landfors, M., Rydén, P. and Hvidsten, T.R., 2011. Classification of microarrays; synergistic effects between normalization, gene selection and machine learning. BMC bioinformatics, 12(1), pp.1-19.
2. Zhao, Z.; Morstatter, F.; Sharma, S.; Alelyani, S.; Anand, A.; Liu, H. Advancing feature selection research. ASU Feature Sel. Repos. 2010, 1–28, doi 10.1.1.642.5862
3. Bolón-Canedo, V.; Sánchez-Marono, N.; Alonso-Betanzos, A.; Benítez, J.M.; Herrera, F. A review of microarray datasets and applied feature selection methods. Inf. Sci. 2014, 282, 111–135.
4. Almugren, N.; Alshamlan, H. A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. IEEE Access 2019, 7, 78533–78548.
5. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature selection: A data perspective. ACM Comput. Surv. (CSUR) 2017, 50, 94
6. Fakoor, R.; Ladhak, F.; Nazi, A.; Huber, M. Using deep learning to enhance cancer diagnosis and classification. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; ACM: New York, NY, USA, 2013; Volume 28.

[1]V. Kalaimani, [2]Dr.R. Umagandhi,

7. Chen, Y.; Li, Y.; Narayan, R.; Subramanian, A.; Xie, X. Gene expression inference with deep learning. Bioinformatics 2016, 32, 1832–1839.

8. Sevakula, R.K.; Singh, V.; Verma, N.K.; Kumar, C.; Cui, Y. Transfer learning for molecular cancer classification using deep neural networks. IEEE/ACM Trans. Comput. Biol. Bioinform. 2018, 16, 2089–2100.

9. Mahfouz, M.A., Shoukry, A. and Ismail, M.A., 2021. EKNN: Ensemble classifier incorporating connectivity and density into kNN with application to cancer diagnosis. Artificial Intelligence in Medicine, 111, p.101985.

10. Vanitha, C.D.A., Devaraj, D. and Venkatesulu, M., 2015. Gene expression data classification using support vector machine and mutual information-based gene selection. procedia computer science, 47, pp.13-21.

11. Sreepada, R.S., Vipsita, S. and Mohapatra, P., 2014, An efficient approach for classification of gene expression microarray data. In 2014 Fourth International Conference of Emerging Applications of Information Technology ,pp. 344-348.

12. Li, Z., Xie, W. and Liu, T., 2018. Efficient feature selection and classification for microarray data. PloS one, 13(8), p.e0202167.

13. Potharaju, S.P.; Sreedevi, M. Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance. Clin. Epidemiol. Glob. Health 2019, 7, 171–176.

14. Zahoor, J. and Zafar, K., 2020. Classification of microarray gene expression data using an infiltration tactics optimization (ITO) algorithm. Genes, 11(7), p.819.

15. Mallick, P.K., Mohapatra, S.K., Chae, G.S. and Mohanty, M.N., 2020. Convergent learning–based model for leukemia classification from gene expression. Personal and Ubiquitous Computing, pp.1-8.

16. Liao, Q.; Ding, Y.; Jiang, Z.L.; Wang, X.; Zhang, C.; Zhang, Q. Multi-task deep convolutional neural network for cancer diagnosis. Neurocomputing 2019, 348, 66–73.

17. Kong, Y. and Yu, T., 2018. A deep neural network model using random forest to extract feature representation for gene expression data classification. Scientific reports, 8(1), pp.1-9.

18. Elbashir, M.K., Ezz, M., Mohammed, M. and Saloum, S.S., 2019. Lightweight convolutional neural network for breast cancer classification using RNA-seq gene expression data. IEEE Access, 7, pp.185338-185348.

19. Sevakula, R.K., Singh, V., Verma, N.K., Kumar, C. and Cui, Y., 2018. Transfer learning for molecular cancer classification using deep neural networks. IEEE/ACM transactions on computational biology and bioinformatics, 16(6), pp.2089-2100.

20. Xia, C., Xiao, Y., Wu, J., Zhao, X. and Li, H., 2019, A convolutional neural network based ensemble method for cancer prediction using DNA methylation data. In Proceedings of the 2019 11th International Conference on Machine Learning and Computing (pp. 191-196).

21. Yang F. and K. Z. Mao, "Improving robustness of gene ranking by multicriterion combination with novel gene importance transformation," Int. J. Data Mining Bioinf., vol. 7, no. 1, pp. 22-37, 2013

22. Yang F.and K. Z. Mao, "Robust feature selection for microarray data based on multicriterion fusion," IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 8, no. 4, pp. 1080-1092, 2011.

23. Wang, G.-G.; Deb, S.; Coelho, L.d.S. Elephant herding optimization. In Proceedings of 2015 3rd International Symposium on Computational and Business Intelligence (ISCBI 2015), Bali, Indonesia, 7– 9 December 2015; pp. 1–5.

24. Umagandhi, R., 2020. Hybrid Ensemble Feature Selection (HEFS) Model for Gene Expression Microarray Data. European Journal of Molecular & Clinical Medicine, 7(3), pp.5022-5036.

25. M. Anthimopoulos et al., Lung pattern classification for interstitial lung diseases using a deep convolutional neural network, IEEE Trans. Med. Imaging 35 (2016), no. 5, 1207–1216.

26. Zhou, Z., Huang, G. and Wang, X., 2019. Ensemble convolutional neural networks for automatic fusion recognition of multi-platform radar emitters. ETRI Journal, 41(6), pp.750-759.

27. Deb, K. and Deb, D., 2014. Analysing mutation schemes for real-parameter genetic algorithms. International Journal of Artificial Intelligence and Soft Computing, 4(1), pp.1-28.