Sequential Feature Maps with LSTM Recurrent Neural Networks for Robust Tumor Classification

C. BATUR ŞAHİN, B. DİRİ

Abstract— **In the field of biomedicine, applications for the identification of biomarkers require a robust gene selection mechanism. To identify the characteristic marker of an observed event, the selection of attributes becomes important. The robustness of gene selection methods affects the detection of biologically meaningful genes in tumor diagnosis. For mapping, a sequential feature long short-term memory (LSTM) network was used with artificial immune recognition systems (AIRS) to remember gene sequences and effectively recall learned sequential patterns. An attempt was made to improve AIRS with LSTM, which is a type of RNNs, to produce discriminative gene subsets for finding biologically meaningful genes in tumor diagnosis. The algorithms were evaluated using six common cancer microarray datasets. By converging to the intrinsic information of the microarray datasets, specific groups such as functions of the co-regulated groups were observed. The results showed that the LSTM-based AIRS model could successfully identify biologically significant genes from the microarray datasets. Furthermore, the predictive genes for biological sequences are important in gene expression microarrays. This study confirmed that different genes could be found in the same pathways. It was also found that the gene subsets selected by the algorithms were involved in important biological pathways. In this manuscript, we tried an LSTM network on our learning problem. We suspected that recurrent neural networks would be a good architecture for making predictions. The results showed that the optimal gene subsets were based on the suggested framework, so they should have rich biomedical interpretability.**

Index Terms— **Biomarker discovery, Deep learning, Gene selection, Robustness, Tumor classification.**

I. INTRODUCTION

MICROARRAY TECHNOLOGY is a technology with a high process capacity that can identify thousands of genes at the same time. It is used to identify disease-related genes by comparing gene expression in a diseased and normal cell. Obtaining pointer genes from high-throughput experiments instead of creating models provides advantages for biomarker discoveries. It can be used in gene expression profiling, the diagnosis of diseases, and pharmacogenetics areas. Some of the genes in gene expression data provide us with important information about disease diagnosis.

Feature selection methods generally influence the performance of biomarker discoveries. Pointer discovery needs a robust feature selection method for microarray datasets. The groups formed by the associated features are generally mentioned as intrinsic specific groups, such as functions, and present in high-dimensional datasets. The current research aimed to develop and assess a method of converging to co-regulated feature groups in microarray datasets, thus, addressing the problem of robust feature groups with high-accuracy classifications. The recent advancements in deep learning techniques in machine learning introduce a strong alternative to high-throughput experiments. The methodology in this research is the immune-based feature selection that is utilized for the discovery of optimal feature sets, enhancing robust tumor classification. The recent research efforts that have utilized feature selection methods include deep learning, feature selection, and classification of cancer microarray datasets. The in-depth investigation of feature selection to enhance the diagnosis of diseases will provide a significant contribution to the literature in the biomarker discovery domain. This study presented and tested a novel method of leveraging feature selection that resulted in the improved classification of tumor diagnosis.

In the study conducted by [1], a new framework of feature selection based on recurrent neural network (RNN) was suggested to select a subset of features. The suggested model was applied to select features from microarray data for cell classification. Feature selection models were implemented based on gated recurrent unit (GRU), long short-term memory (LSTM), RNN and bidirectional LSTM for microarray datasets. In the study carried out by [2], a deep neural network model was improved by feature selection algorithms in predicting various biomedical phenotypes. Five binary classification methylome datasets were selected to compute the prediction performances of CNN/DBN/RNN models by utilizing the feature selected by the eleven feature selection algorithms. The results showed that the Deep Belief Network (DBN) model

CANAN BATUR ŞAHIN is with the Department of Computer Engineering, University of Malatya Turgut Ozal, Malatya, Turkey (e-mail: canan.batur@ozal.edu.tr) https://orcid.org/0000-0002-2131-6368

BANU DİRİ is with the Department of Computer Engineering, Yildiz Technical University, Istanbul (e-mail:diri@ce.yildiz.edu.tr).
https://orcid.org/0000-0002-6652-4339

Manuscript received August 10, 2019; accepted September 24, 2020.
DOI: 10.17694/bajece.604885

utilizing the features selected by SVM-RFE usually had the best prediction accuracy on the five methylome datasets.

In the research conducted by [3], a novel approach was established based on clustering-centered feature selection for the classification of gene expression datasets. According to the experiments, the suggested feature clustering support vector machine (FCSVM) was capable of achieving efficient performance on gene expression datasets.

In the research performed by [4], the most recent studies using deep learning to establish models for cancer prognosis prediction were reviewed. This study revealed that the application of deep learning in cancer prognosis was equivalent to or better in comparison with the current approaches.

In the study conducted by [5], a convolutional neural network (CNN) deep learning algorithm was investigated for the classification of microarray datasets. The promising results proved that CNN had superiorities in terms of accuracy and minimizing gene in classifying cancer.

In the research conducted by [6], the performance of deep neural networks for the classification of gene expression microarrays was analyzed. The experimental results suggested that deep learning needs high-throughput datasets to achieve the best performance.

In the study carried out by [7], deep learning-based algorithms were developed to make a tumor diagnosis, reveal biomarkers and genetic changes, pathological features. An overview showed that deep learning-based approaches for pathology gave promising results for big data.

The recent developments of robust biomarker techniques with deep learning introduce a strong alternative to tumor diagnosis. The principal contribution of the current manuscript is presented below:

1. The current study presents an extensive framework in which the learning and combination of features are carried out in a novel way, by establishing a biomarker discovery model.
2. The Artificial Immune Recognition (AIRS) algorithm was trained by deep learning-based approaches to sequentially learn biologically meaningful genes in order to predict a tumor diagnosis.
3. We leveraged the LSTM deep learning technique to capture the long context correlations in a cancer microarray dataset.
4. We optimized deep-learned features by utilizing LSTM recurrent neural networks (RNNs) to detect co-regulated specific groups, such as functions, in high-dimensional data.
5. We examined the possibility of utilizing Deep Neural Network (DNN) recurrent neural network models to learn disease-related genes, and then we used them for the prediction of important biological pathways.

In this study, LSTM-based AIRS version 1, LSTM-based AIRS parallel version 1, LSTM-based AIRS version 2, and LSTM-based AIRS parallel version 2 algorithms were developed to discover optimal biological gene sequences. The suggested algorithms were compared with the traditional genetic algorithm and genetic algorithm-based artificial immune systems. All the experiments in this study used the microarray dataset.

The present research is structured in the following way. Section II summarizes the feature subset group. Sections III briefly explains the systems. Section IV describes the methodology and framework. Sections V and VI focus on the results and performance analysis, and the conclusions.

II. Feature Subset Group

Robust tumor classification was constructed with an ensemble gene selection framework. The framework uses feature subset groups comprised of the associated attributes. Feature groups are created using group formation algorithms that run separately on sub-samples of the training dataset. The bootstrap method was used to ensure the stability of training samples in the presence of variations. The associated feature groups were created with filter-based feature selection methods.

Density-based feature groups were created by kernel density estimation, which was calculated using equation (1). The kernel function is determined by the C_{j+1} formula to identify the consecutive locations of the kernel function. Kernel density estimates were made to locate dense feature groups, and then the most relevant groups were selected.

$$C_{j+1=} \frac{\sum_{i=1}^{k} fi\; K(\frac{C_j - fi}{h})}{\sum_{i=1}^{k} K(\frac{C_j - fi}{h})} ,\; j = 1,2, ... \qquad (1)$$

In Eq. (1), variables h, k, f_i, and K represent the kernel bandwidth, the nearest neighbor number, the number of attributes in the dataset, any attribute that is represented by parameter f_i, and the kernel function, respectively.

The usefulness of attribute subsets in the CFG was identified by Eq. (2). The intuitive usability of a subset of S was based on the heuristic evaluation function.

$$\mathrm{merit_S} = \frac{\mathrm{k * rcf}}{\sqrt{\mathrm{k + (k-1) * rff}}} \qquad (2)$$

Variables k, rcf, and rff represent the number of attributes, the mean attribute-class correlation, and the correlation between the mean attributes, respectively.

In Eq. (3), the information gain function identifies the significance of a given attribute in the full feature set. The entropy criteria were used to determine feature knowledge.

$$E = -\sum_{i=1}^{M} (f_t(i) \log(f_t(i)) \qquad (3)$$

Parameters f_t and M represent any attribute and the data numbers, respectively.

III. Systems

A. Long Short-Term Memory (LSTM)

LSTM is a variation of RNN architecture and one of the most effective solutions to sequence prediction problems because of the recognition of patterns in data sequences. Since LSTM possesses a certain type of memory, it can selectively remember patterns for a long time. It is quite a reasonable approach to predict the period with the unknown duration between important events. Fig. 1 demonstrates the architecture of the LSTM recurrent neural network. It comprises a self-recurrent connection and three gates, input, forget, and output, which are responsible for remembering things and manipulating the memory cell. Interactions between the memory cell and its environment are modulated by the mentioned gates. The input gate is responsible for adding information to the cell state. The forget gate allows the cell to remember or forget the cell's previous state. The output gate selects beneficial information from the current cell state and shows it out.

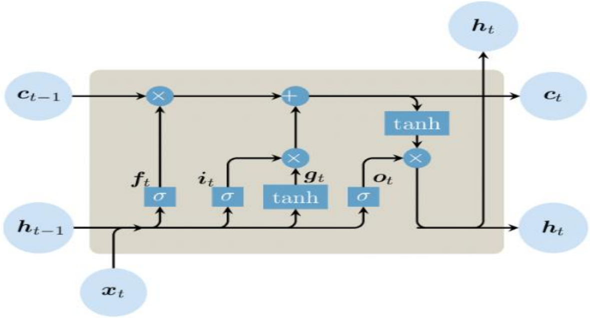

Fig. 1. Architecture of the LSTM Recurrent Neural Network [8].

Each LSTM block has input and output gates that learn to activate or deactivate to obtain new information, change the cell state, and activate it to affect other cells and network outputs. X(t) is an input for the antigenic pattern at time t. For each time series, one LSTM block changes the output of the new cell state (Ĉ) at time t, which acts as the current cell state at time t. A tanh layer is added to Ĉ(t), which represents the new state of the cell at time t. Then, the old cell state C(t-1) is updated as C(t). The modulation and output gates are represented by g(t) and O(t), respectively.

B. Artificial Immune Recognition System (AIRS)

The artificial immune recognition system (AIRS) represents an intelligent system that is inspired by the natural immune system.

AIRS depends on the stages of initialization, memory cell recognition, resource competition, and the selection of memory cells. The normalization of the dataset is performed at the initialization stage in the range of [0, 1]. Then, the affinity threshold is computed using Eq. (4). The affinity threshold represents the average affinity between antigens in the training set. In Eq. (4), variables *n, ag*$_i$, and *ag*$_j$ refer to the number of antigens in the dataset, any antigen, and the next antigen in the dataset, respectively. Antigens are trained in the artificial recognition ball (ARB) pool during the resource competition stage. The stimulation value is assigned to every ARB to compete for limited resources. Memory cells are selected at the end of the resource competition stage. The evolved memory cell pool indicates the quality of the classification process. The affinity between two antigens in the training set is calculated using Eq. (5). Eq. (6) calculates stimulation. If a memory unit and an antigen have the same class label, this refers to the stimulation value in affinity. If they have a different class label, this refers to the stimulation value in the Euclidean distance. In Eq. (6), m$_c$ represents the memory cell.

$$Affinity\ threshold = \sum\nolimits_{i=1}^{n}\sum\nolimits_{j=i+1}^{n}\left(\frac{\mathrm{Affinity}\left(\mathrm{ag_i}\,,\mathrm{ag_j}\,\right)}{\frac{\mathrm{n(n+1)}}{2}}\right) \qquad (4)$$

$$Affinity\left(ag_i\,,ag_j\,\right) = 1 - Eucliden\ distance\left(ag_i\,,ag_j\,\right) \qquad (5)$$

$$Stimulation = \begin{cases} Affinity(m_c, ag_i)\ \ if\ m_c.class = \ agi.class \\ 1 - Affinity \qquad\quad otherwise \end{cases} \qquad (6)$$

The first version of the artificial immune recognition system (AIRS1) utilizes the ARB pool as a permanent resource, and the mutation rate is determined by the user. The second version, AIRS2, utilizes the ARB pool as a temporary resource. Therefore, the complexity of AIRS2 is less, and somatic hypermutation is used, which means that the mutation rate is proportional to affinity. The parallel versions of AIRS are PAIRS1 and PAIRS2. The training datasets are separated into np number processes. The AIRS algorithm was run separately on each process and merged with the np memory pool.

C. Genetic Algorithm

The genetic algorithm (GA) represents a stochastic search model and optimization technique that mimics natural evolutionary mechanisms. GA is a population-based algorithm that evolves solutions on the basis of the principles of Darwinism. Each candidate solution is represented by the chromosome and has a fitness value indicating the quality of the solution to a problem. GA starts by generating a random population. Fitness-based selection determines the recombined parent chromosomes in the mating pool. Through crossover and mutation operators, offspring are produced for the next generation. The evolution of successive generations continues until the stopping criterion is achieved. At the final stage, the best solution to a problem is determined.

D. Genetic Artificial Neural Network with the Genetic Algorithm (ANN +GA)

The artificial neural network (ANN) represents a computational structure that models the neural structure of the human brain. Artificial neurons are the basic units, which are connected to weighted values, synapses. The structure comprises input, hidden, and output layers. The input layer provides data to ANN. The hidden layer consists of units transforming input into something in the output layer. The feature subsets of ANN were created with DGF, CFG, and IGFG feature subset groups. The genetic algorithm (GA) was employed to estimate the best input parameters of ANN to train networks.

IV. Methodology and Suggested Framework

The methodology was inspired by a theoretical model of the natural immune system, which describes the functioning and behavior of the immune system and has been an inspiration for a new artificial immune system (AIS).

The theoretical model hypothesizes that “a kind of internal restimulation keeps immune memory preserved for a long time” [9]. To model such an internal restimulation mechanism, this study used an LSTM recurrent neural network. The proposed framework was designed for the robust computation of long-sequence learning. The adaptive immune system is capable of remembering the same antigenic patterns over different periods. An associative immune memory was developed to remember gene sequences as robust patterns. This study developed a mechanism for sequence modeling in which biologically significant gene sequences could be effectively memorized. The immune memory of AIRS was developed based on this methodology to understand the “remember” behavior of the artificial immune system response. The underlying principle of the LSTM-based AIRS is to allow for the preservation of the subpopulation of surviving ARBs as long-lived unit cells. In the LSTM systems, values, for which durations are random and delays between significant events are unknown, can be remembered. The evolution of each ARB was performed with the LSTM block for long time series. All recognition cells were remodeled with LSTM gates during the training of the system and then treated with the metadynamics of AIRS. LSTM evolves sub-populations of memory cells and treats them as network inputs. The proposed framework formulates long-sequence learning problems with LSTM memory blocks, as shown below:

The output of the network h(t) is computed by utilizing the formula presented below. C_t^j refers to the memory amount of every jth LSTM unit at time t.

$$h_t^j = \sigma_t^j \tanh\left(c_t^j\right) \tag{7}$$

σ(t) denotes the output gate in which the memory content exposure is managed.

The output gate is expressed by the following equation:

$$\sigma_t^j = \sigma(W_0 X_t + U_0 h_{t-1} + V_0 c_t)^j \tag{8}$$

σ represents the standard sigmoid function, while V0 represents a diagonal matrix. Ĉ(t) denotes a new memory content of the memory unit, which is updated by partially forgetting the current memory and adding new memory contents to c(t).

$$c_t^j = f_t^j c_{t-1}^j + i_t^j \hat{C}_t^j \tag{9}$$

The novel memory contents are presented below:

$$\hat{C}_t^j = \tanh(W_c X_t + U_c h_{t-1})^j \tag{10}$$

The current memory forgetting gate is modulated by f(t). Input gate i(t) modulates the addition degree of new memory content to the memory cell. V_f and V_i are diagonal matrices [10].

$$f_t^j = \sigma(W_f X_t + U_f h_{t-1} + V_f c_{t-1})^j \tag{11}$$

$$i_t^j = \sigma(W_i X_t + i\, h_{t-1} + V_i c_{t-1})^j \tag{12}$$

Fig. 2, Fig. 3, and Fig. 4 demonstrate the pseudocode of LSTM-AIRS, the general flowchart of the suggested model, and the framework of the LSTM-AIRS architecture, respectively.

Procedure: LSTM-AIRS

```
Step 1: {InitializeAntibodyPool (AntibodySet)}
Step 2: {InitializeFeatureSet(Ω)}
Step 3: [Train] {1...N} (Input Size)
Step 4: {IntroduceAntibodyPool} (Ab$_N$)
Step 4: FOR I← Iteration Number DO
    Step 5: Affinities←{calcAffinities (Ab$_i$)}
    Step 6: Clone$_{num}$ ← Select{Affinity (Ab$_i$)}
    Step 7: Fitness ←Accuracy (Feature set (Ω), Ab$_i$)
    Step 8: Antibody$_{LSTM}$ ← CreateLSTMMemory (Ab$_i$, t, State_id)
    Step 9: FOR (Ab$_i$ ∈ Clone$_{num}$)
        Step 10: ClonesAntibodies ←(CloneandHypermutated (Abi))
        Step 11: CloneAffinities←{calcAffinities (Abi)}
        Step 12: Fitness←Accuracy (Feature set (Ω), Abi)
        Step 13: {UpdateAntibodyPool} ←(CloneAffinities)
        Step 14: Ĉ$_{LSTM}$ ←UpdateLSTMMemory (Antibody$_{DNN}$, t, State_id)
        Step 15: State_id ←State_id+1
        Step 16: C$_{LSTM}$ ←UpdateLSTMMemory (Ĉ$_{LSTM}$, t,State_id)
    Step 17: END FOR
    Step 18: bestAffinity←{getBestAffinity (CloneAffinities)}
    Step 20: Ω*← NewFeatureSet(C$_{LSTM}$, bestAffinity)
Step 20: END FOR
```

Fig.2. Pseudocode of LSTM-AIRS

LSTM represents a variation of RNN cells, which is easier to train when the vanishing gradient problem is avoided. The vanishing gradient problem emerges during the training of RNNs with long sequential time series data, and the gradient of error concerning the model parameters at early time steps approaches zero. This indicates that it becomes more challenging for the model to learn long-term dependencies in the input time series. For each time series, the propagation of inputs occurs through the recurrent neural network with the memory cells that are newly calculated.

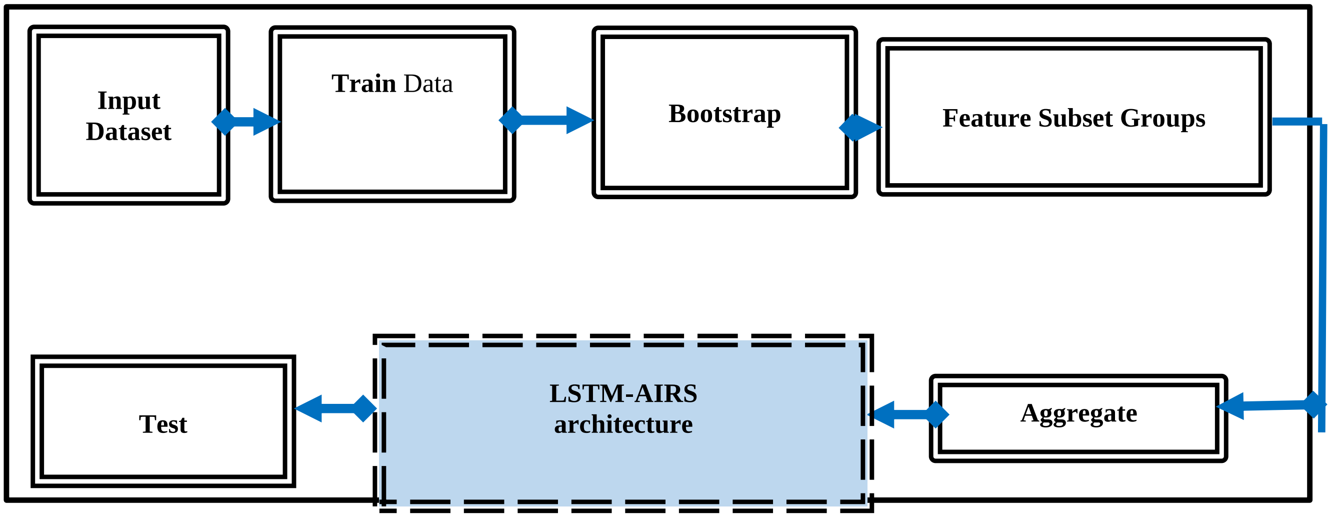

Fig. 3. General flowchart of the suggested model

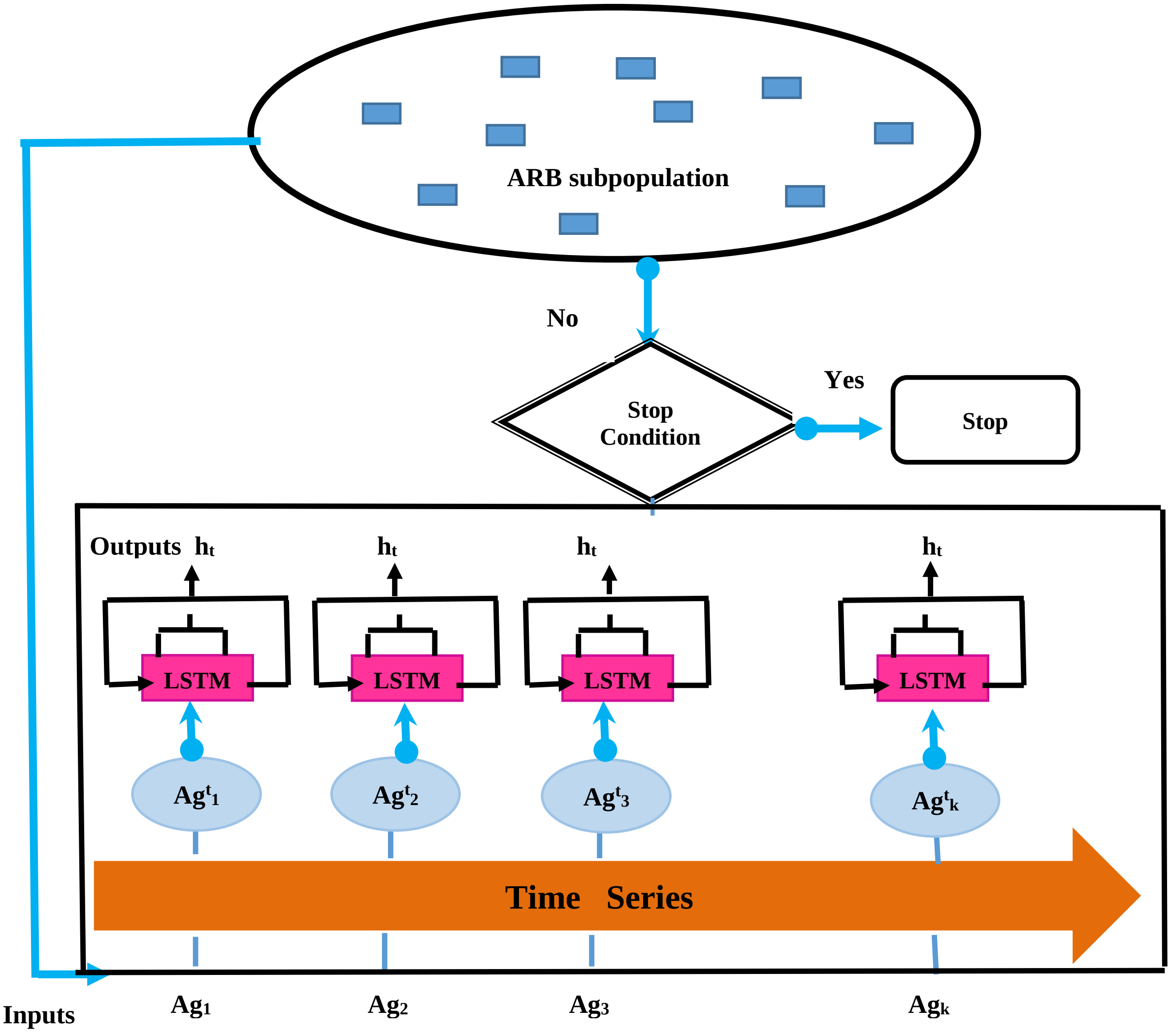

Fig. 4. Pseudocode of LSTM-AIRS

IV. Results and Performance Analysis

This study focused on the robustness of gene selection algorithms. Robust gene sequences were evaluated by the accuracy of each class computed using equation (13):

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{13}$$

The characteristics of gene expression profile (GEP) datasets may cause over-fitting and bias selection problems if small gene sets are selected from large dimensional features. There may be a chance of finding high classification performance from small gene subsets in high-dimensional datasets. Therefore, robust gene selection algorithms are required for GEP datasets [11]. Some predictor genes were reported in this part. The DAVID and REACTOME programs were used for the biological knowledge discovery of the selected genes. The biological information was extracted from the UNIPROT and NCBI ENTREZ databases. The k-NN and SVM classifiers were used directly as a classifier to measure the classification accuracy of optimal gene subsets. Ten-fold cross-validation was utilized to evaluate the classification model. It was aimed to find reliable accuracy on the training set and test set separately. The frequency measure was used as a measure of the significance of gene subsets in long sequences.

TABLE I
TUMOR-RELATED GENES FOR COLON, LUNG, AND PROSTATE DATASETS

Datasets		***Colon Dataset***				***Lung Dataset***				***Prostate Dataset***			
Algorithms	**Group Type**	***SVM 10 CV on***		***KNN 10 CV on***		***SVM 10 CV on***		***KNN 10 CV on***		***SVM 10 CV on***		***KNN 10 CV on***	
		Training Set	***Test Set***	***Training Set***	***Test Set***	***Training Set***	***Test Set***	***Training Set***	***Test Set***	***Training Set***	***Test Set***	***Training Set***	***Test Set***
LSTM-AIRS1	DGF	79.5	82.1	72.7	76.9	93.2	94.4	96.6	97.2	81.5	80.9	79.1	80.9
	CFG	75.4	79.4	71.4	76.4	82.1	86.1	80.8	88.8	71.9	70.7	77.4	79.8
	IGFG	72.5	76.9	74.3	80.7	84.4	88.1	82.6	83.3	68.4	69.6	75	77.1
LSTM-PAIRS1	DGF	75.4	79.4	71.4	75.4	**98.4**	**98.6**	**97.2**	**97.9**	70.4	72.1	74.1	83.9
	CFG	75.5	80.7	76.3	75.5	86.2	92.8	84.8	93.3	71.1	71.5	77.9	85.7
	IGFG	67.9	71.3	69.7	67.9	87.5	90.7	84.6	91.6	68.9	71	74.6	88.1
LSTM-AIRS2	DGF	81.6	84.6	75.5	79.2	90.7	91.6	89.8	91.6	75.3	76.1	79.1	80.9
	CFG	75.2	76.9	71.4	76.9	82	95.8	82.2	91.3	70.2	71.4	65.8	66.7
	IGFG	67.3	69.2	60	69.2	83.1	86.1	86.8	88.8	70	71.4	60.1	61.9
LSTM-PAIRS2	DGF	**89**	**92.3**	**85.3**	**91.2**	98.3	92.8	97.2	91.6	**92.1**	**84.4**	**88.4**	**76.7**
	CFG	80	86.2	58.7	90.5	82.7	91.6	80.6	91.6	89.5	84.4	87	79
	IGFG	71.4	68.6	71.2	74.5	88.9	97.2	87.5	94.4	72.9	71.5	69.7	70
GA	DGF	63.2	61.5	71.4	76.9	86.1	93.7	90.3	91.6	52.3	80.2	57.1	71.6
	CFG	63.2	69.2	61.2	67.8	82	86.1	78.6	90.8	38	51.8	46.9	61.9
	IGFG	62.7	66	60.8	65.3	82.7	88.8	81.3	89.7	61.7	61.9	59.7	63.5
ANN+GA	DGF	83.6	85.7	81.6	86	90	94.3	91.5	97.9	79.5	80.9	85.1	90
	CFG	71.4	82.6	75.3	87.1	89.2	93	79.3	91.7	65.4	71.5	59.2	79.2
	IGFG	68.9	70.9	66.9	69.8	83.4	95.5	91	91.6	69.3	70.2	53.1	71.6

In all tables, we marked in bold the jointly selected genes based on DGF, CFG, and IGFG for the tested methods. The selected gene sequences were analyzed based on the gene function and pathway analysis. For the colon dataset, the LSTM_PAIRS2 algorithm exhibited the best performance with the training accuracy of 89% and predicting accuracy of 92.3% using the SVM classifier and the training accuracy of 85.3% and predicting accuracy of 91.2% using the k-NN classifier based on DGF. {R28608, T94993, L19437 (TALDO1), M82919 (GABRB3), T55780} is the selected gene subset in the colon dataset based on DGF. The results in Table 1 show important tumor-related genes preserved in long-lived unit cells. TP53 regulates the transcription of cell cycle genes and performs transcriptional regulation by TP53 pathways [12]. P53 plays an important role in colorectal cancers, and p21 is a critical effector of butyrate-induced growth arrest in colonic

TABLE II
TUMOR-RELATED GENES FOR SRBCT, LYMPHOMA, AND LEUKEMIA DATASETS

Datasets		***SRBCT Dataset***				***Lymphoma Dataset***				***Leukemia Dataset***			
Algorithms	**Group Type**	***SVM 10 CV on***		***KNN 10 CV on***		***SVM 10 CV on***		***KNN 10 CV on***		***SVM 10 CV on***		***KNN 10 CV on***	
		Training Set	***Test Set***	***Training Set***	***Test Set***	***Training Set***	***Test Set***	***Training Set***	***Test Set***	***Training Set***	***Test Set***	***Training Set***	***Test Set***
LSTM-AIRS1	DGF	80	81.5	75.6	78.9	81.7	82.6	81.6	94.4	73.3	83.2	78.7	87.7
	CFG	71.2	73	78.5	79.7	70.3	78.5	80.3	83.2	76	82	80	85
	IGFG	60	61.1	63.4	65.7	65.3	70	77.5	78.3	84.8	85.9	82.4	86.6
LSTM-PAIRS1	DGF	74.2	76.9	76.9	80	84.6	93.3	78.8	95.9	84.6	86.9	80	91.2
	CFG	72	73.8	70	79.6	72.4	78.3	73.4	76.9	77.1	82.5	79	87.3
	IGFG	59.8	68.4	56.7	67.3	70.2	81.6	70	72.7	76.2	86.6	79.8	85
LSTM-AIRS2	DGF	80	92.3	82.6	84.6	**91.8**	**92.3**	**93.8**	**94.6**	84.7	89.2	84.7	93.3
	CFG	74.2	75.4	72.1	76.4	80	84.1	77.2	82.7	77.1	86.6	73.9	93.3
	IGFG	61.3	65	67.3	69.1	81.4	84.6	77.3	92.3	72.4	86.6	68.9	71.9
LSTM-PAIRS2	DGF	**90.8**	**92**	**91.7**	**94.3**	**94.5**	**95.1**	**90.8**	**92.3**	**96.3**	**86.6**	**98.2**	**85.2**
	CFG	76.2	75	71	74.3	86.9	86.8	81.6	84.6	83.3	86.6	85.5	80
	IGFG	65	68	64.3	65.8	80.4	85.9	81.6	92.3	80.5	86.6	69.2	72.3
GA	DGF	66	76.9	68	69.2	81.6	84.6	83.6	92.3	86.6	90.2	80	85.9
	CFG	46	58	58	69.2	72.3	79.2	71.4	78.5	77.9	85.7	63.7	66.7
	IGFG	45.7	46.1	48	49.8	75.5	78.1	73.4	76.9	77.1	83.8	71.9	79.8
ANN+GA	DGF	78.5	82	74.3	86	83.3	90.8	89.7	96	**92.9**	**93.7**	**91.1**	**92.9**
	CFG	65.8	78.7	76	82.3	77.1	82.4	79.5	92.3	79.3	89	70.9	82.1
	IGFG	58.2	65.4	52	69.2	76.3	81.5	77.1	79.8	80	82.8	75.3	81.4

TABLE III
EXPERIMENT RESULTS FOR MICROARRAY DATASETS.

Classifiers	Group Type	***Colon Dataset***	***Lung Dataset***	***Prostate Dataset***	***SRBCT Dataset***	***Lymphoma Dataset***	***Leukemia Dataset***
MLP	DGF	79.2	83.4	72.1	70.6	82.7	83.5
	CFG	76.6	82.6	70.8	69.6	82.4	82.7
	IGFG	74.3	80.5	71.3	68.4*	80.6	81.5
LSTM	DGF	**84.7**	**89.6**	75.7	**77.6**	88.3	85.3
	CFG	82.1	87.4	74.3	75.3	87.8	85.3
	IGFG	77.9	84.5	72.4	73.8	85.7	78.5
GRU	DGF	82.6	88.2	**76.8**	78.1	**89.7**	**88.2**
	CFG	83.7	85.6	71.6	76.7	83.6	84
	IGFG	75.6	81.4	70.3	76.3	83.8	76.1

cancer cells. Activating transcription factor 4 (ATF4) is effective in colorectal cancer [13].

For the lung dataset, the LSTM_PAIRS1 algorithm exhibited the best performance with the training accuracy of 98.4% and predicting accuracy of 98.6% using the SVM classifier and training accuracy of 97.2% and predicting accuracy of 97.9% using the k-NN classifier with selection in the lung dataset based on DGF. The USP32P1, CD44, HCRTR2, TNFSF4, NUP98, CCNO, NCF2, and TCEB3-AS1 genes commonly involved in the metabolism of RNA and class I MHC mediated antigen processing and presentation pathways [14] are expressed in lung cancer.

In the prostate dataset, the highest classification performance was achieved through the LSTM_PAIRS2 algorithm with the optimal gene set consisting of eight genes, LYL1 (39971_at), HSD17B3 (39978_at), BMPR2 (39998_at), SCARF1 (40034_r_at), ASIC2 (40317_at), DYNC1I1 (40319_at), HSPA4L (40354_at), and ITGA2B (40643_at)}, with the training accuracy of 92.1% and predicting accuracy of 84.4% using the SVM classifier and training accuracy of 88.4% and predicting accuracy of 76.7% using the k-NN classifier based on DGF. The SLC35A2 gene participates in seven different pathways. The ACD gene is involved in eleven different biological pathways, such as reproduction, meiosis, meiotic synapsis, and cell cycle pathways. The RPL10A gene is involved in 32 different biological pathways. The RPL10A gene is involved in SRP-dependent cotranslational protein targeting to membrane pathway and plays a role in proteins upregulated in prostate cancer. The RPL10A gene takes part in the base regulation of expression of SLITs and ROBOs. SLITs and ROBOs are expressed during the development of prostate cancer.

For the SRBCT dataset, the LSTM_PAIRS2 algorithm exhibited the best classification performance with the {GNAO1, DNAJA1, PSMB10, PRKCE6(PML), MDK, XRCC5} gene subset with the training accuracy of 90.8% and predicting accuracy of 92% using the SVM classifier and training accuracy of 91.7% and predicting accuracy of 94.3% using the k-NN classifier based on DGF. The selected gene, PSMB10, participates in the stabilization of p53, regulation of TP53 expression, p53-dependent G1 DNA damage response, and p53-dependent G1/S DNA damage checkpoint pathways. XRCC5 is X-ray repair cross complementing 5, ERCC2 is ERCC excision repair 2, TFIIH core complex helicase subunit [15], POLG is DNA polymerase gamma, catalytic subunit, EIF2S1 is eukaryotic translation initiation factor 2 subunit alpha.

For the lymphoma dataset, the results show that the LSTM_AIRS2 algorithm exhibited the best performance with the training accuracy of 91.8% and predicting accuracy of 92.3% using the SVM classifier and training accuracy of 93.8% and predicting accuracy of 94.6% using the k-NN classifier. The LSTM_PAIRS2 algorithm exhibited the classification performance with the training accuracy of 94.5% and predicting accuracy of 95.1% using the SVM classifier and training accuracy of 90.8% and predicting accuracy of 92.3% using the k-NN classifier while selecting the {GENE595X, CARP cardiac ankyrin repeat protein, GENE585X, TNNT1 troponin T1, skeletal, slow.GENE771X, Homo sapiens mRNA; cDNA } gene subset in the lymphoma dataset based on DGF. Type II transmembrane protein contains C-lectin domains and is related to DC-SIGN [16].

For the leukemia dataset the {ATP6V0C; ATPase, H+ transporting, lysosomal 16kDa, V0 subunit c, CTSD; cathepsin D lysosomal aspartyl peptidase, AKT1; v-akt murine thymoma viral oncogene homolog 1, CSRP1; cysteine and glycine-rich protein 1, TGFBI; transforming growth factor, beta-induced, 68kDa, CCND3; cyclin D3, SERPINB1; serpin peptidase inhibitor, clade B ovalbumin, member 1} gene subset was selected based on DGF by the LSTM_PAIRS2 algorithm with the training accuracy of 96.3% and predicting accuracy of 86.6% using the SVM classifier and training accuracy of 98.2% and predicting accuracy of 85.2% using the k-NN classifier.

Table 3 shows the experimental performance results of Multi-Layer Perceptron (MLP), Long-Short-Term Memory (LSTM), Gated Recurrent Unit (GRU) classifiers using Deep Learning 4j (DL4J). The results present the GRU and LSTM classifiers had best performances by 89.7 and 89.6 accuracies with Lymphoma and Lung datasets respectively based on the DGF feature group. Also, GRU classifier had accuracy 88.2 with Leukemia dataset based on the DGF feature group. The results of Table 3 presents the worst classification accuracy obtained by MLP classifier with SRBCT dataset with 68.4 accuracies based on the IGFG feature group.

Figure 5-7 show the classification performance of the algorithms on the basis of the datasets. The results showed the lung dataset obtained the best classification accuracy of 94.2% by the LSTM-PAIRS1 and ANN+GA algorithms using the k-NN classifier. At the same time, the prostate dataset obtained the lowest classification accuracy of 50.6% by the GA algorithm using the SVM-train and SVM classifier.

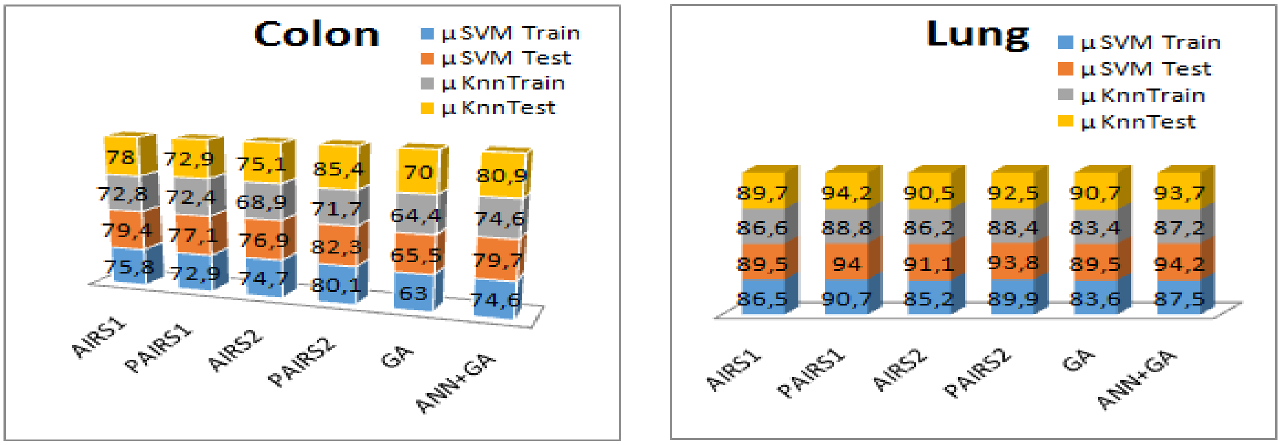

Fig. 5. Classification performance of the colon and lung datasets based on the algorithms

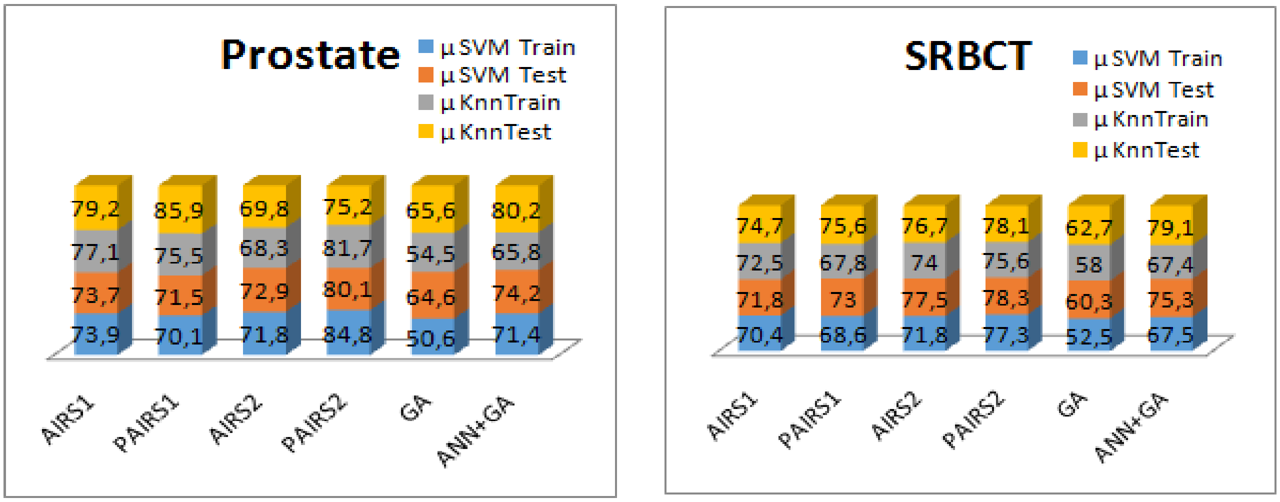

Fig. 6. Classification performance of the prostate and SRBCT datasets based on the algorithms

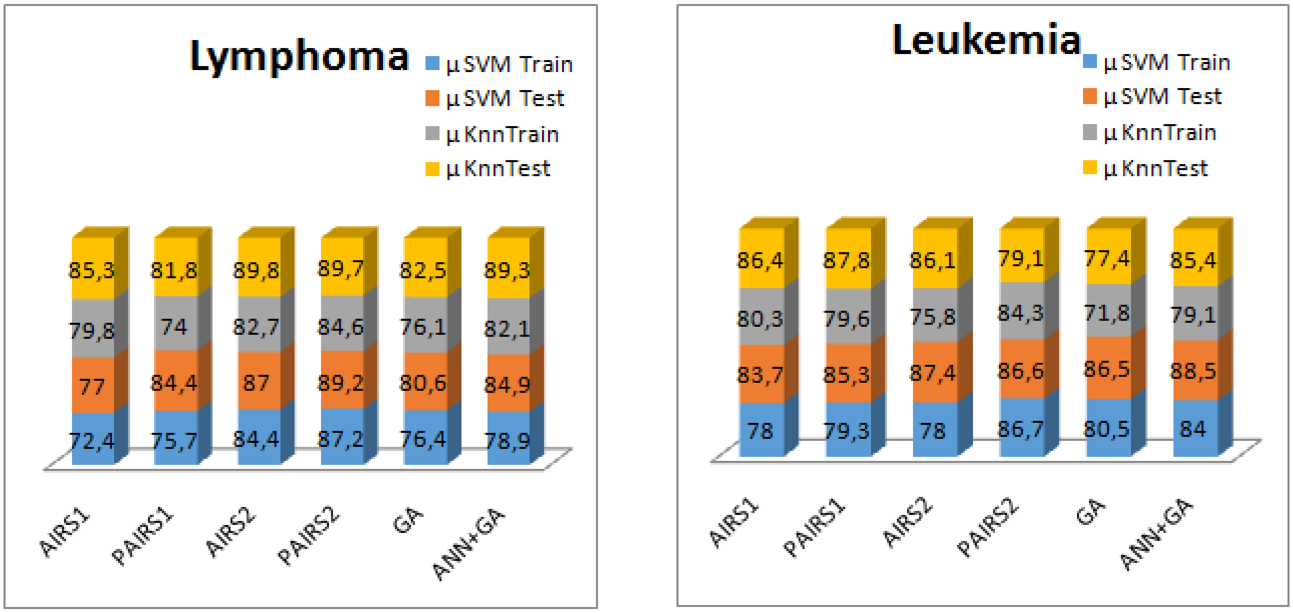

Fig. 7. Classification performance of the lymphoma and leukemia datasets based on the algorithms

V. Conclusion

The framework showed that predictive genes for biological sequences were important in gene expression microarrays. It was also found that the gene subsets selected by the algorithms were involved in important biological pathways. LSTM was used to learn sequences over time. The suggested framework was proposed for converting immune memory into an intelligent network system. The analysis of gene sequences was performed, and informative genes from each dataset were detected. This study confirmed that different genes could be found in the same pathways. Optimal gene subsets were obtained from six commonly used microarray datasets.

In future research, our aim is to investigate different deep neural network models (e.g., the BiLSTM network, CNN network) to improve the performance of the proposed model. Furthermore, it is crucial to evaluate the proposed methodology on other datasets. For future studies, we also aim to conduct an assessment of biomarker detection on different types of techniques presented for the detection of Coronavirus (COVID-19).

References

[1] Chowdhury S., Dong X., Li X.,“Recurrent Neural Network Based Feature Selection for High Dimensional and Low Sample Size Micro-array Data”, IEEE International Conference on Big Data (Big Data), 978-1-7281-0858-2/19/$31.00, 2019.

[2] Chen Z., Pang M., Zhao Z., et al. “Feature Selection may Improve Deep Neural Network for the Bioinformatic Problem”, Bioinformatics. Doi: 10.1093/bioinformatics/btz763, 36(5), 2019.

[3] Mabu A., Prasad R., Yadav R., "Gene Expression Dataset Classification Using Artificial Neural Network and Clustering-Based Feature Selection", International Journal of Swarm Intelligence Research. Doi: 10.4018/IJSR20200104, 2020.

[4] Zhu W.., Xie L., Han J., Guo X., "The Application of Deep Learning in Cancer Diagnosis", Cancers. *12*(3), 603; https://doi.org/10.3390/cancers12030603, 2020.

[5] Zeebaree D., "Gene Selection and Classification of Microarray Data Using Convolutional Neural Network", International Conference on Advanced Science and
Engineering (ICOASE), Doi: 10.1109/ICOASE.2018.8548836, 2018.

[6] Nava A. R., Sánchez J. S., Alejo R., Flores-Fuentes A. A., Rendón-Lara E., "The Application of Deep Learning in Cancer Diagnosis", Pattern Recognition, Doi:10.1007/978-3-319-92198-3_11, 2018.

[7] Jiang Y., Yang. M., Wang S., Li X.., Sun Y., "Emerging Role of Deep Learning-Based Artificial Intelligence in Tumor Pathology", Cancer Communications, doi: 10.1002/cac212012, 2020.

[8] Rubvurm M., Körner M., "Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images", The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), USA, 2017.

[9] Dasgupta D., Nino L.F., Immunological computation: theory and applications. 2009 by Taylor & Francis Group, LLC. Book, 2009.

[10] Chung Y., et al., "Empirical evaluation of gated recurrent neural networks on sequence modeling", Neural ComputAppl arxiv:1412.3555v1, 2014.

[11] Luque-Baena R.M., Urda D., Gonzalo Claros M., Franco L., and Jerez J.M., "Robust genesignatustes from microarray data using genetic algorithms enriched with biological pathway keywords", Journal of Biomedical Informatics, http://dx.doi.org/10.1016/j.jbi.2014.01.006, 2014.

[12] https://machinelearning-blog.com/2018/02/21/recurrent-neural-networks/

[13] https://www.cs.waikato.ac.nz/ml/weka/

[14] https://worldwidescience.org/topicpages/i/immunity-related+gtpase+irgml.html

[15] https://www.genecards.org/cgiin/carddisp.pl?gene=PCBP4

[16] http://www.bionewsonline.com/

[17] Loscalzo S., YU L., Ding C., "Consensus Group Stable Feature Selection", June 28–July 1, Paris, France, 2009.

[18] Loscalzo S., YU L., Ding C., "Stable Feature Selection via Dense Feature Groups", August 24–27, Las Vegas, Nevada, USA, 2008.

[19] Farhan M., Mohsin M., HamdanA., Bakar A.A., "An evaluation of feature selection technique for dentrite cell algorithm", IEEE, 2014.

[20] Schmidhuber, J., Wierstra, D., and Gomez, F. J. Evolino: "Hybrid neuroevolution/optimal linear search for sequence prediction". In Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI), pp. 853–858, 2005

[21] Ferreira A.j. (2014), "Feature selection and discretization for high-dimensional data", Phd Thesis.

[22] Mazel J. (2011), "Unsupervised network anomaly detection", Phd Thesis.

[23] Polat K., Güneş S., Sahan S., Kodaz H., (2005), "A New Classification Method for Breast Cancer Diagnosis: Feature Selection Artificial Immune Recognition System (FS-AIRS)", Berlin.

[24] Menon A. edited by. (2004). Frontiers of Evolutionary Computation, Kluwer academic Publishers, Pittsburgh, USA.

[25] Oh S., Lee J.S., Moon B.R. (2004). Hybrid Genetic Algorithms for Feature Selection.Vol. 26, No.11, November.

[26] Brownlee J. (2005). "Clonal Selection Theory & Clonalg the Clonal Selection Classification Algorithm (CSCA), Technical Report.

[27] Dudek G. (2012). "An Artificial System for Classification with Local Feature Selection", IEEE Transaction on Evolutionary Computation. December 2012, Czestochowa.

[28] Wang K., Chen K., Adrian A., (2014), "An improved artificial immune recognition system with the opposite sign test for feature selection", Knowledge-Based Systems, Taiwan.

[29] Daga M., Lakhwani K., (2013), "A Novel Content Based Image Retrieval Implemented By NSA of AIS", International Journal of Scientific & Technology Research.

[30] Farhan M., Mohsin M., Hamdan A., Bakar A.A., (2014), "An Evaluation of Feature Selection Technique for Dendrite Cell Algorithm", IEEE.

[31] Gu F., Greensmith J., Aickelin U., (2008), "Further Exploration of the Dendritic Cell Algorithm: Antigen Multiplier and Time Windows", ICARIS 2008.

[32] Wollmer M., Eyben F., Rigoll G., (2008), "Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening", Journal of Selected Topics in Signal Processing.

[33] Daga M., Lakhwani K., (2013), "A Novel Content Based Image Retrieval Implemented By NSA Of AIS", International journal of scientific & technology research

[34] Kalousis A., Prados J. and Hilario M., (2007), "Stability of Feature Selection Algorithms: a study on high-dimensional spaces", Knowledge and Information Systems.

[35] Saeys Y., Abeel T., and Peer Y.V. (2008), "Robust Feature Selection Using Ensemble Feature Selection Techniques". In Proceedings of the ECML Conference, pages 313-325.

BIOGRAPHIES

Canan BATUR ŞAHİN is an Assist. Prof. Dr. at Malatya Turgut Ozal University. She received her diploma and Ph.D. degrees in Computer Engineering from Yildiz Technical University. Her research interests include optimization, artificial intelligence, and software engineering.

Banu DİRİ is a professor at Yildiz Technical University, and she works on natural language processing. She authored more than 150 publications in this field. Her research interests include speech recognition, natural language processing, and machine learning.