EAI Endorsed Transactions
on Pervasive Health and Technology

Research Article **EAI.EU**

Design of a Novel Ensemble Model of Classification Technique for Gene-Expression Data of Lung Cancer with Modified Genetic Algorithm

Prem Kumar Chandrakar[1,*], Akhilesh Kumar Shrivas[2], Neelam Sahu[3]

[1]Department of Computer Science, Mahant Laxminarayan Das College, Raipur (C.G.) India.
[2]Department of Computer Science and Information Technology, Guru Ghasidas Vishwavidyalaya, Bilaspur. India.
[3]Department of Information Technology and Computer Science, Dr. C.V. Raman University, Kota, Bilaspur. India.

Abstract

INTRODUCTION: Gene expression levels are important for identifying and diagnosing diseases like cancer. Gene expression microarray information contains a high extent feature set, which minimizes the performance and the accuracy of classifiers.
OBJECTIVES: This paper proposes a Modified Genetic Algorithm (MGA) that is based on Classifier Subset Evaluators – Genetic Search (Eval-CSE_GS) for selecting the relevant feature subsets. The MGA feature selection procedure is applied to microarray information for cancer patients that minimize a high dimension feature subset into low dimension feature subsets.
METHODS: The various data mining methods for classifying the various kinds of cancer disease patients are presented. The proposed model refers to an ensemble model (PEM) for the organization of cancer disease by reducing the feature subsets, which results show improvements in the success rate.
RESULTS: The proposed ensemble model obtains the accuracy of 94.58%, 96.56% and 97.04% for PEM-1 to PEM-3, respectively.
CONCLUSION: Our proposed MGA-PEM model gives satisfactory results for cancer identification and classification.

Received on 27 August 2020; accepted on 12 December 2020; published on 08 January 2021

Keywords: Gene Expression, Modified Genetic Algorithm (MGA), Ensemble, Proposed Ensemble Model (PEM), Microarray, Lung Cancer.

doi:10.4108/eai.8-1-2021.167845

1. Introduction

Gene-expression patterns are attributes of disorder diagnosis, which can be applied to accurately classify cancer. Nowadays, many data mining and classification strategies like Naive Bayes and J-48 are being developed in the research community, in which most of the methods are applied to cancer disorder data and its organization [1, 2]. This supported organic phenomenon amounts from microarray information and gene-disease relationships may be detected using machine-learning algorithms and owing to the high dimensionality of microarray information data sets, which is often challenged with over-fitting, poor performance, and low potency. Given these challenges, there are some discrimination methods for the classification of tumors proposed by Dudoit [3] through the use of high-density DNA sequences and oligonucleotides. Feature choice [4], ensemble call trees [5], and ensemble neural networks [6] also appear to be effective and possible solutions. Although many researchers have explored cancer classification, few of them have centered on the combinatory ensemble methodology with a support vector machine or are inconclusive in terms of the classifiers' performance. This paper proposes a Modified Genetic-Algorithm (MGA) and Projected Ensemble Model (PEM) as a learning algorithmic rule that remarkably improves the accuracy and strength of

[*] Corresponding author. Email: prem.k.chandrakar@gmail.com

the classification model. The proposed MGA-PEM is a combination of feature choice and an ensemble classifier that will avail additional sample data, having higher accuracy in comparison with existing classifiers. Also, we explored how the various feature choice strategies affect the performance of the classifiers and the way several options ought to be selected to urge the most effective use of the classifiers. Finally, the proposed method is analyzed and compared with existent work models. This paper is divided into the following sections: Section 2 reports related work in the field whereas the proposed materials are covered in Section 3. Section 4 presents the performance measures, and the pseudocode is explained in section 5. In section 6, results are analyzed. The paper ends with the discussion of results in section 7.

2. Related Work

The existing analysis of information mining of biomedical datasets within the literature is quite extensive. For example, the author Sathyadevi [7], used Classification and Regression Trees CART, C4.5, and Iterative Dichotomiser ID3 algorithms to diagnose effectively hepatitis disease. In the same vein, CART calculation performed best in identifying the disease.

Roslina and Noraziah [8] used Support Vector Machines for classification and prognosis estimation of hepatitis, where they utilized a wrapper method, to remove noise in data before determining classification. The mix wrapper based most of the strategies and support vector machines on sensitive classification results.

Huang et al. [9] presented a filter-based technique that selected two parameters (age and number of claims) to attain a similar prediction authenticity. The comparative analysis of this technique suggests that the proposed technique is a feasible and efficient technique to reduce the size of the healthcare databases.

Larranaga et al., [10] reviewed machine-learning techniques for bioinformatics. In this review, the authors aimed to compare different machine-learning approaches for bioinformatics like clustering, supervised-classification, models for knowledge-discovery, etc. In this review, the different applications of current machine-learning techniques for bioinformatics like systems biology, genomics, proteomics, evolution, and text mining are also presented.

Inza et al. [11], used DNA microarray datasets related to the determination of cancer, including cancer and leukemia. The outcomes featured that filter and wrapper based mostly on quality determination approaches prompt extensively better exactness regarding correlation with the non-gene choice system, combined with intriguing and hanging spatiality reductions.

Cancer is one of the most threats worldwide According to the World Health Organization – WHO, 9.6 million malignancy-associated diseases were accounted for in 2018 [12].

The natural mensuration procedure is dependent upon Ribonucleic Acid, and RNA, rather than supermolecular particles. This owns to the fact that the RNA arrangements layout hybridizes with their reciprocal RNA or polymer group while this property needs supermolecules. Indeed, the multivariate are novel advances for gene delivery, containing an enormous scope of qualities (in thousands) and a low number of examinations (in handfuls).

In AI classification, datasets normally include many measurements and testing iterations. The point of the cistron decision is to look out for a lot of qualities that best segregate natural examples of different sorts. The chosen qualities are biomarkers, acting as a "marker board" for examination. Although an information gain may be observed in this ‘marker board’ rank [13], there is also an entropy on this model-based data. Multivariate Gaussian generative models were, therefore, used to model the data with variable ordinary distributions.

Rajagopal, Kundapur, & Hareesha has proposed an ensemble approach using the concept of stacking for effective network intrusion detection. The varying gene expression can be efficiently analyzed using microarray where all the genes of a particular organism are placed in different grooves on a slide. Gene expression data could be effectively maintained and processed using statistical methods to analyze diseases much easier [14].

The state of a cell communicated by the layout of RNA will thusly serve to be of great help to check whether a cell might be a normal or a variation one [15].

The use of machine learning in cancer diagnosis is becoming more feasible as algorithms become less prone to error and noise, and as the volume of training data increases[16].

The proposed SVM and KNN methods are tested and the accuracy of both the approaches are recorded as 71.52% and 94.74%, respectively. The essential idea of the Genetic Algorithm is utilized to generate solutions and to determine improvement issues [17].

Zahoor and Zafar [18] have discussed the microarray technology that produces thousands of genes in a single study or record. Sampling shortages, digital errors, and cursing microarray data are some of the difficulties to accurately detect cancer cells and to avoid overdoses. They have shown that, apart from the data, the accuracy and reliability of the model are equally different and, therefore, both factors should be considered when evaluating the model. Both multiple voters and soft ensembles produced similar results.

Table 1. Correctly and incorrectly classification [20]

Classifier	Correctly classified Instances	In Correctly classified Instances	Kappa statisctics
Naive Bayes	72%	28%	0.29
Logistic Regression	69%	31%	0.19
K-NN	72%	28%	0.24
Random Forest	70%	30%	0.173

Chen, Meng, and Su [19] have discussed the Gene selection algorithm for small data editing problems. Well-chosen genetic selection of the algorithm should select a set of genes that achieve the highest performance and size, and for this, the genetic set should be as small as possible. Many gene selection algorithms are available but suffer from a low performance or large size. Collective genetic selection is a proposed Algorithm, WERFE, which belongs to the wrapper method inside an RFE framework and maintains a combination of genetic selection cross-certification. The comparative analysis of different lung cancer prediction based on correctly and incorrectly classification is presented in Table 1.

3. Proposed Model and Materials

Here, the proposed architectures are divided into four phases: In Phase 1, the clinical trial proposes a Modify Genetic Algorithm (MGA) as a feature-selection and dimensional-reduction method to cut back the feature set. Phase 2 explores the partition of optimized information set into coaching and testing using 10-fold cross-validation. In Phase 3, the clinical test explored the model building of various people and projected ensemble models, and, lastly, Phase 4 is employed for model validation and comparison, and an ensemble model is projected. In this paper, the general method of the experimental work, shown in Figure 1, embodies four phases:

1. **Phase 1** explores the projected feature selection technique wherever we acquired the gene expression information, integrating microarray information from lung cancer patients. Then, we preprocessed and applied normalization technique for information smoothness, before additional analysis. After that, features were selected applying a MGA feature reduction technique.

2. **Phase 2** describes the partition of information into coaching and testing 10-fold cross-validation. The partition of information is an incredibly important step for the development of data processing techniques.

3. **Phase 3** describes the model building using various data mining techniques. In the model building process, we have supplied the training data set into various individuals and ensemble techniques and train the model.

4. **Phase 4** discussed the model validation and comparison of the performance of various people and ensemble models. The performance measures assess the strength of the model that classifies the lung carcinomas with high accuracy.

3.1. Dataset Description

This work utilized two microarray datasets of quality articulations from completely different groups. These two datasets have various qualities (one can be linearly separated and the other one non-linearly separated). The essential data set was from disease patients with 2 variations of the lung cancer (myeloid lung- AML and lymphoblastic lung-ALL) (Golub TR). Data has 2 subsets (Number of samples is 203 and number of attributes is 12600): the instructing set is utilized to choose qualities and change loads of the classifiers; an independent check set is utilized to appraise the presentation of the classifier.

Gene expression lung cancer dataset contains 203 snap-solidified respiratory organ tumors (n=186) (where n lies between 0 to 4 as presented in Table 1) and normal or normal adjacent to tumor lung samples (n=17). The total informational collection contains 125 glandular malignant growth samples that are identified with clinical data and with infinitesimal life systems slides from neighboring segments.

The carcinoma dataset of 203 specimens (Dataset A) includes lung adenocarcinoma (n=127), squamous cell lung carcinoma (n=21), carcinoids (n=20), small-cell lung carcinoma (SCLC) (n=6), and normal lung (n=17) samples.

Other adenocarcinoma samples (n=12) were suspected to be extrapulmonary metastases dependent on clinical history (See the sample dataset from SampleData.xls, which is printed as supporting data on the PNAS registering machine, source is www.pnas.org). The dataset incorporates exclusively adenocarcinomas and normal lung samples. The following cryptography of

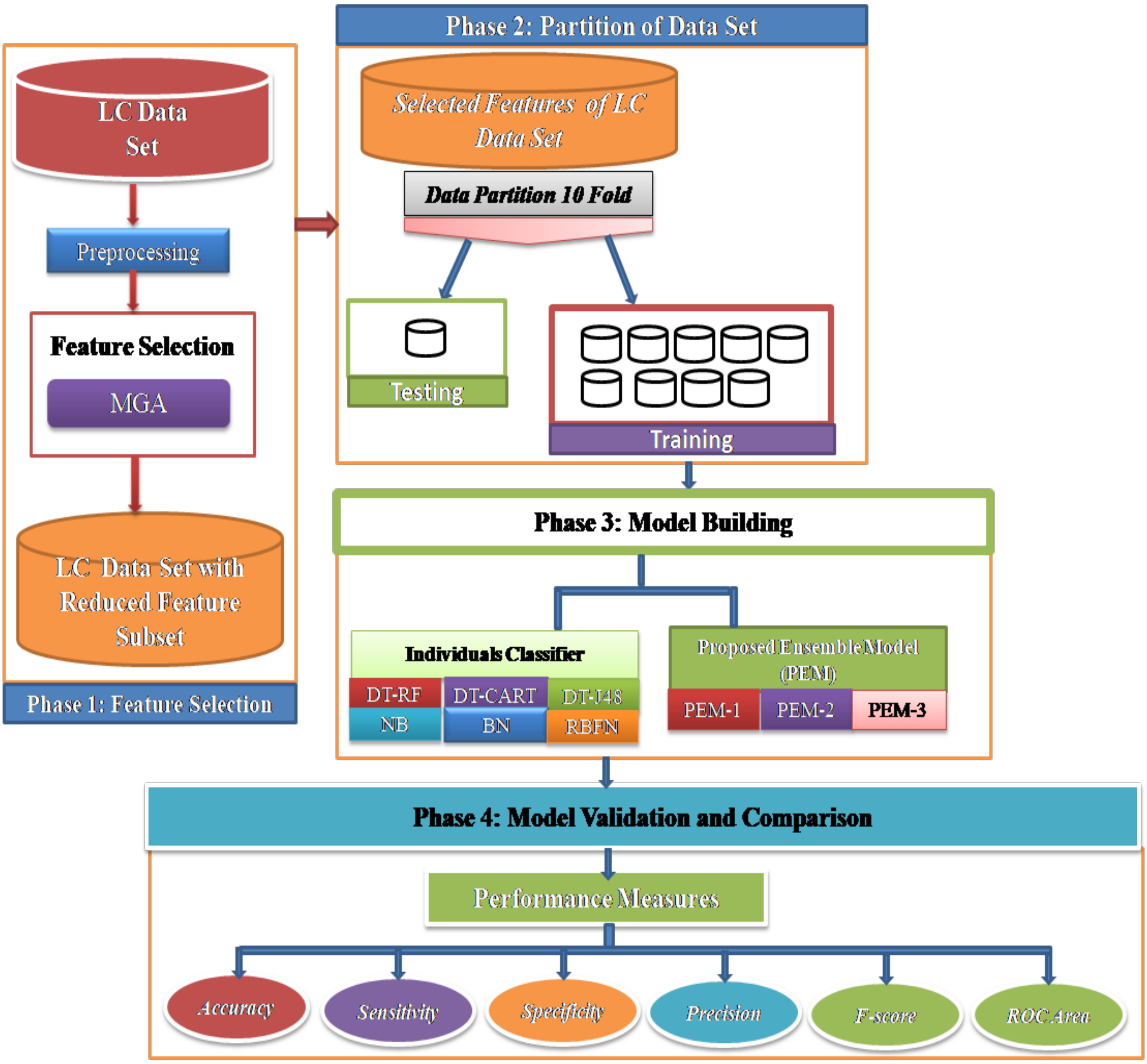


Figure 1. The architecture of the proposed model.

diagnostis classes was used, which is presented in Table 2.

In this study we have use, lung cancer gene expression datasets, which were collected from biomedical data repository, USA. The dataset of 203 samples encompasses 12,600 features (genes). The data matrix of the gene expression data is presented in Table 3. The acronyms used in Figure 1 are as follows:
LC = Lung Cancer, MGA = Modified Genetic Algorithm, DT-RF = Decision Tree-Random Forest, DT-CART= Decision Tree-Classification and Regression Tree, DT-J48 = Decision Tree-J48, NB= Navie Bayes, BN= Bayes

Table 2. Lung Cancer dataset value description

Adeno by adenocarcinoma	0
Normal	1
Squamous by cell lung carcinoma	2
Compensation for Occupational Injuries and Diseases (COID)	3
Secondary Maximum Contaminant Level (SMCL)	4

Table 3. Data matrix for lung cancer dataset.

Sample/Genes	gene1	gene2	....	Gene12600
Sample 1	.....	.....	.....	.....
Sample 2	.....	.....	.....	.....
....	.....	.....	.....	.....
....	.....	.....	.....	.....
Sample 203	.....	.....	.....	.....

Network, RBFN= Radial Basis Function Network, and ROC-Area= Reciver operating charactecteristcs-Area.

3.2. Feature Selection

Feature choice is an optimization technique that is used to remove the unsuitable feature set from the original feature house and improves the classification accuracy exploitation relevant or a necessary feature set. This research work used Genetic algorithmic rule (GA) or Genetic Search (GS) as options selection.

To reduce features from the dataset, we also used Principal Component Analysis (PCA) [21] for dimensional reduction, and feature selection techniques [22] to reduce the features from the original feature space.

Genetic-Algorithms (GA) [17, 23] were utilized to the order of development to determine advancement issues. The most cited presentation of the one of a kind Genetic algorithmic standard was developed by John Holland who described it in the mid-1970s [24] and hereditary calculations is versatile inquiry procedures that bolstered the standards of a normal activity in science. They utilized a populace of competitor arrangements changed after some time to unite and relate the ideal answer quickly, i.e. the appropriate response house is looked at equal that helps in keep from local optima. For highlighting a decision, there is an answer here and there and a firm length double string embodying a list of features - the value of each position inside the string “speaks” to the nearness of non-attendance of a choice to include. The algorithmic guideline could be a reiteration technique in any place each age is made by hereditary administrators, for example, the current age individuals by executing hybrid transformation.

Mutations randomly change certain values (thus adding or removing features in this way) in a subset. Crossover combines different attributes with a pair of subsets in a new subset [25].

The use of hereditary administrators in a populace of individuals is chosen by their wellness (how keen a list of capabilities is comparable to an investigative technique). Indeed, higher element subsets are an opportunity to be picked to shape a new set through a hybrid model or transformation. In this way, good subsets are “developed” over time. A fully expanded subset is where all possible local changes have been considered.

3.3. Cross-Validation

Cross-validation could be a statistical procedure to estimate the potential of machine learning models. It is usually employed in applied machine learning to compare and choose a model for a given prognostic modelling downside and as a result it is easier to implement and lower the effect of bias.

3.4. Data mining based mostly classification technique

Classification is one of the necessary data processing applications and its method of classifying the samples into distinguishes information categories. Classification is supervised learning that consists of 2 phases: training and testing. In the training section, a classifier trained exploitation testing information set and trained model tested exploitation testing information set. There are four classification techniques employed in this analysis work for classification of carcinoma genes expression information set [26].

Decision Tree. A decision tree is a choice help apparatus that utilizes a tree-like model of choices and their feasible results, including chance occasion results, asset costs, and utility. it’s a method to show a calculation that exclusively contains contingent administration explanations. Call trees are generally used in tasks research, explicitly in choice examination to help to delineate a strategy to potentially achieve an objective. Moreover, they are a popular apparatus in Artificial Intelligence.

Random Forest. Random Forest (or RF) [18, 27] is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. Random Forests are often used when we have very large training datasets and a very large number of input variables (hundreds or even thousands of input variables). A random forest model is typically made up of tens or hundreds of decision trees.

CART (Classification and Regression Trees). CART is a non-prohibitive Decision Tree (DT) learning strategy that builds either order or relapse trees, contingent upon whetherthe subordinate variable is all out or numeric. It builds a double DT by partitioning the record at each hub, in sync with a performance of one property.

CART uses the Gini index for determining the best divide statistical techniques encompass usually used in health care in support of the classification of various diseases.

J48. J48 can be an approach uses to deliver call tree created by Ross Quinlan. J48 is the related extension of Quinlan's previous ID3 calculation. the decision trees made by J48 will be sent to contributory for arrangement, and for this method of reasoning, J48 is regularly said as an applied science classifier [28].

Naive Bayes (NB). Naive Bayes classifiers are a group of direct "probabilistic classifiers" in light of applying Bayes' hypothesis with solid (guileless) freedom suspicions between the highlights. they're among the least difficult Bayesian system models [29].

Bayes Network (BN). Technique based on machine learning, which was introduced by Judea Pearl in 1985 it referred to as theorem Network [30]. The Bayesian network is proficient and effective to property for representing and calculation under the situation of vagueness [31, 32]. Their achievement has shown the way to up to date furry of methods for learning Bayesian networks from data.

Radial Basis Function Network (RBFN). The planning of a supervised neural network will be followed. Considering the design of a neural network [33] as a curve fitting (approximation) downside in the high-dimensional house within the RBF neural network, there is a completely different definite approach [34] that presents the satisfactory suit for coaching statistics with the criteria of "best fit" measures.

4. Performance Measures

Different execution estimates, for example, exactness, affectability, particularity, accuracy, and F-measure that are determined with the help of the disarray framework. Disarray framework incorporates boundaries like genuine True_Positive (TP), True_Negative (TN), False_Positive (FP) and False_Negative (FN). The confusion matrix for 2 classes is shown in Table 4, in which TP alludes assortment of positive examples that are appropriately ordered by classifier. By contrast, True Negatives (TN) and False Positives (FP) are an assortment of negative examples. False Negatives (FN) are relative to the measure of positive examples that are inaccurately grouped.

On the off chance that the whole assortment of cases is N, at that point based on most part show in Table 3 the following Confusion-matrix applied arithmetic execution estimates will be assessed. Arrangement precision by Confusion-matrix quantifies the extent of the right forecast thinking about the positive and negative sources of information. It is noteworthy the dataset dispersion, which may bring the wrong ends in regard to the framework execution represented in Equation (1).

Table 4. Confusion-matrix for positive and negative examples

Actual vs. predicted	Positive	Negative
Positive	True_ positive (TP)	False_ negative (FN)
Negative	False_ positive (FP)	True_ negative (TN)

$$Accuracy = (TP + TN)/N \tag{1}$$

Sensitivity. Sensitivity measures the extent of true positives, i.e. the adaptability of the framework on anticipating the best possible qualities of the cases provided, being represented in Equation (2).

$$Sensitivity = TP/(TP + TN) \tag{2}$$

Specificity. Specificity measures the extent of true negatives, i.e. the adaptability of the framework on foreseeing the best possible qualities for the cases that are the option of the ideal one, being determined through Equation (3).

$$Specificity = TN/(TN + FP) \tag{3}$$

It is the pace of occasions that are characterized appropriately among the after-effects of the classifier, being represented in Equation (4).

$$Precision = TP/(TP + FP) \tag{4}$$

F-measure. The mean of exactitude and recall.

$$F - measure = 2x((precision\_recall)/(precision\_recall)) \tag{5}$$

Development of an efficient classifier.

F-Score. The F1 score is determined by the following equation:

$$2 * ((precision * recall)/(precision + recall)) \tag{6}$$

It is conjointly known as the F Score or the F Measure. The F-Score may be, therefore, determined by the following equation:

$$F\_Score = 2 * ((precision * recall)/(precision + recall)) \tag{7}$$

ROC Area. In a Receiver Operating Characteristic (ROC) curve calculation the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold [35].

5. Pseudocode of Proposed Work

The proposed algorithm is a Modified Genetic Algorithm based of feature selection techniques. The Pseudo code is divided into four basic sections to explore the research work.

1. The first section is based on feature selection that means that our first algorithm is a modified algorithm by attribute section that is entitled 'Pseudo-Code Main'.

2. The second section is a main modified algorithm based on the Genetic Algorithm entitled Pseudo-Code of Modified Genetic Algorithms.

3. The third section of pseudo code contains Pseudo-Code of Ensemble Model that is based on a genetic based ensemble algorithm, which provides the classification accuracy with a gene expression dataset.

4. Section four is part of an ensemble algorithm that is entitled Pseudocode of Sub-function Model. The modified algorithm based on different classifiers like BN, NB, RBFN, RF and CART.

Pseudocode-1: Pseudocode Main. Pseudocode 1 shows the proposed Modified Genetic Algorithms, i.e. another form existing search-based feature selection techniques as simple Genetic Algorithms.

```
Start
$Set \leftarrow fuupper - limit$;
$Set \leftarrow fllower - limit$;
while $count = 1ton$ do
    Call GA_M;
    if ($features >= fl\&\&features <= fu$ )then
    break  else
    count++  end Repeat of GA_M
end
Returns GA_M (GA_FLC with Subset features)
End
```

Pseudocode-2: Pseudocode of Modified GA . In this section, we have proposed a pseudo code of feature selection. The proposed algorithm is used to reduce the number of features of a dataset, very easily and in less time.

```
Input: Lung Cancer (LC) dataset features;
Output: find the feature subset;
Generate the initial population ;
------------------------------------------------------------
Start
Function GA_M ($Set \leftarrow Feature$)
{
Search Methods
Genetic Search: Performs a search using the
  simple genetic algorithm ;
$Set \leftarrow crossover$;
$Set \leftarrow maxGen$;
$Set \leftarrow mutation$;
$Set \leftarrow population$;
$Set \leftarrow reportFrequency$;
$Set \leftarrow seed$;
}
Compute Fitness Function;
Fitness function ();
{
Attribute Evaluators
CfsSubsetEval:individual predictive ability ;
locallyPredictive =TRUE;
missingSeparate=fALLSE;
}
UNTIL the population has converged
End
```

Pseudocode-3: Pseudocode of Ensemble Model . In this section, we have proposed a pseudo code of Modified Genetic Algorithms. For it, we have used a Genetic Algorithm, Fitness Function and Attribute Evaluators Classification with the selected data set. The proposed algorithm can enable us to classify the dataset easily in a short time. Whereas Pseudo-code 2 shows the proposed Ensemble Model, the following Pseudocode 3 is a subpart of Pseudocode 2. In other words, the Pseudocode 2 shows how to ensemble the different learning classifiers and how they work, and the Pseudocode is the group of classifiers, i.e. used in Pseudocode 2.

Pseudocode-4: Pseudocode of Sub-function Model . In this section, we have proposed a Pseudo-code of an Ensemble Model. We have combined two or more classifiers and find out the classification performance. Based on the proposed algorithm, we can also optimize the dataset. The pseudocode is divided into the following 3 parts:

```
Input: Input- features (GA_FLC Subset dataset
  (LC)). The total Selected features of GA;
Output: Acc=Accuracy;
________________________________________
Start
Function Ensmble_Model (Classifier accuracy)
{
Folds I = cut (seq(1,nrow(GA_FLC
  dataset)),breaks=10)
while $i = 1 to 10$ do
    if $I$ <= 9 then
    call subfunction # FOR TRAINIG SET
    else
    call= subfunction # FOR TESTING
end
}
ACC=Find Accuracy (GA_FLC dataset
  Algorithms)
Return Ensmble_Model (ACC)
End
```

```
Part (a)
Start
Sub-function ()
{
Algorithms = [NB, DT-J48, DT-RF] ;
Classifiers= matrix(row_length=len(target),
  column_length=len(algorithms));
Or
i, algorithm in enumerate(algorithms);
{
classifiers [i] = algorithm.fit(train,
  target).predict(test);
}
}
End
```

```
Part (b)
Start
Sub-function ()
{
Algorithms = [NB, DT-J48, DT-RF] ;
Classifiers= matrix(row_length=len(target),
  column_length=len(algorithms));
Or
i, algorithm in enumerate(algorithms);
{
classifiers [i] = algorithm.fit(train,
  target).predict(test);
}
}
End
```

```
Part (c)
Start
Sub-function ()
{
Algorithms = [BN, RBFN] ;
Classifiers= matrix(row_length=len(target),
  column_length=len(algorithms));
Or
i, algorithm in enumerate(algorithms);
{
classifiers [i] = algorithm.fit(train,
  target).predict(test);
}
}
End
```

6. Results & Analysis

The main objective of this research work is to reduce optimal number of features of a dataset and achieve maximum performance with our proposed model.
Trial work is done using the Waikato Environment for Knowledge Analysis - WEKA open-source information mining programming for Windows. WEKA constitutes an Artificial Intelligence tool that facilitates information preprocessing and arrangement.
The proficiency of the genetic search CSEGS calculation is checked with various classifications of datasets (see Table 5). Figure 2 shows the feature subset of data sets. Firstly, by reducing features 5 times at the same procedure when we found to reduce new data sets again reduce at last to 46 features then stops the feature and data reduction process.
Secondly, the accuracy of individual classifiers with MGA FST was found out and, thirdly, the accuracy of the proposed MGA-PEM model was also determined.

Table 5. LC data set with reduced features subset using MGA FST

Data Set	No. of Features	No. of Instances
LC-1CSEGS	5669	203
LC-2CSEGS	935	203
LC-3CSEGS	170	203
LC-4CSEGS	64	203
LC-5CSEGS	46	203

The last reduce data set is with 46 number of feature and again can't be reduce the features of dataset due to stop the data reduction process.
Second by finding out the accuracy of individual classifiers with MGA FST and third by finding out the accuracy of proposed MGA-PEM model.

In the fourth step, performance measures of the best-proposed model with LC-5CSEGS Dataset in the fifth step is relative to the Comparison of classification accuracy of the proposed and existing feature selection techniques. At last, the comparative analysis of the proposed model with different existing techniques is performed. As shown in Table 4 and illustrated in Figure 2, the LC dataset with reduced features subset uses MGA FST. After the fifth reduction, data reduction is stopped and in the sixth step, no more data can be reduced.
In the second step of experiments, the accuracy of individual classifiers with MGA FST of all five data sets are reduced. Data reduction can also perform better accuracy results in comparison with the first reduction dataset.

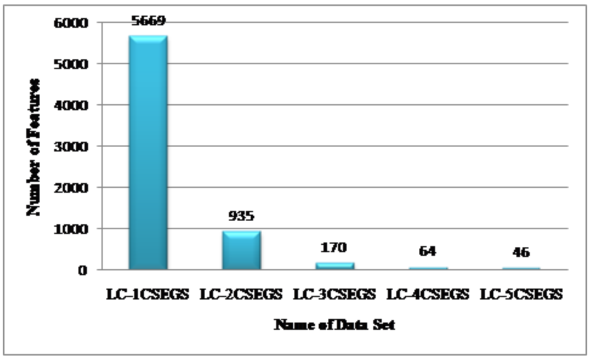


Figure 2. Feature subset of data sets.

Similarly, the fourth dataset is reduced and the last fifth dataset reduction perform the best accuracy results of different models. Figure 2 shows the accuracy of individual classifiers with different feature subsets and Table 6 shows the accuracy of individual classifiers, using MGA FST.

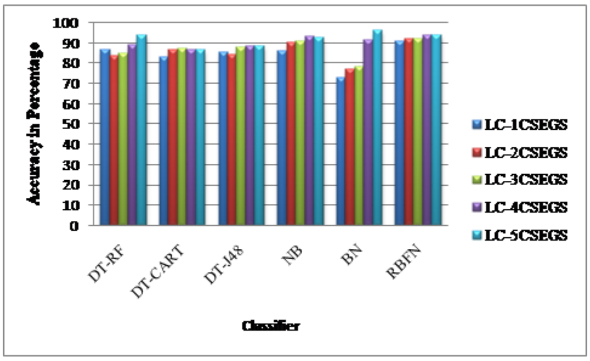


Figure 3. Accuracy of individual classifiers with different feature subset.

As shown in Table 5, the fifth dataset reduction BN and RBFN are performing the best results in

Table 6. Accuracy of individual classifiers with MGA FST

Individual Model	DT-RF	DT-CART	DT-J48	NB	BN	RBFN
Data Set						
LC-1CSEGS	86.69%	83.25%	85.71%	86.20%	72.90%	91.13%
LC-2CSEGS	83.74%	86.69%	84.23%	90.14%	76.84%	92.11%
LC-3CSEGS	84.72%	87.19%	88.17%	91.13%	78.32%	92.11%
LC-4CSEGS	89.16%	86.69%	88.67%	93.10%	91.62%	94.08%
LC-5CSEGS	93.59%	86.69%	88.67%	92.61%	96.05%	94.08%

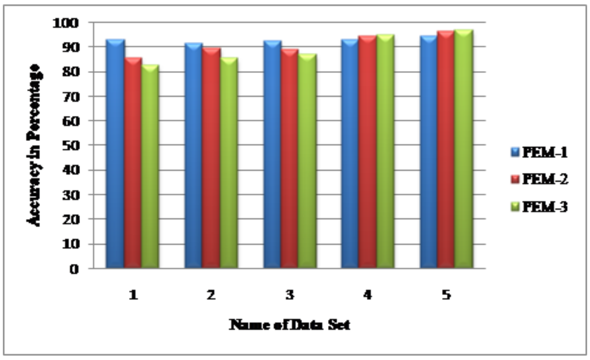


Figure 4. Accuracy of the proposed model with different feature subset

comparison with other models presented in Figure 3.
During the third step of the experiment, three ensemble models were proposed to assess the accuracy of the proposed MFA-PEM model. Table 6 shows the accuracy results proposed in the MGA-PEM model.
From the three proposed models, PEM-3 was shown to perform the best results in comparison with PEM-1 and PEM-2 (Table 7). Figure 4 shows the accuracy of the proposed of the proposed model with a different feature subset.

Table 7. Accuracy of proposed MGA-PEM model

Ensemble Model	PEM1	PEM2	PEM3
Data Set			
LC-1CSEGS	93.10%	85.71%	82.75%
LC-2CSEGS	91.62%	89.65%	85.71%
LC-3CSEGS	92.61%	89.16%	87.19%
LC-4CSEGS	93.10%	94.58%	95.07%
LC-5CSEGS	**94.58%**	**96.56%**	**97.04%**

Finally, the fourth step of the experiment relied on the Performance Measures of Best Proposed Model with LC-5CSEGS data set, followed by the analysis of accuracy, sensitivity (TPR), specificity (1-FPR), Precision, F-Score and the ROC area.
Table 8 and Figure 5 shows the Performance Measures of the Best Proposed Model with LC-5CSEGS Data Set.

The fifth step of the experiment.

In this study, the genetic algorithm was used to reduce the minimum number of the features of the gene expression data set and the accuracy when classifying the data set.
The proposed model was compared with a PCA algorithm and CSEBF feature selection technique. Then, the proposed ensemble model is compiled to give better performance with reduced features subset by the proposed modified genetic algorithm.

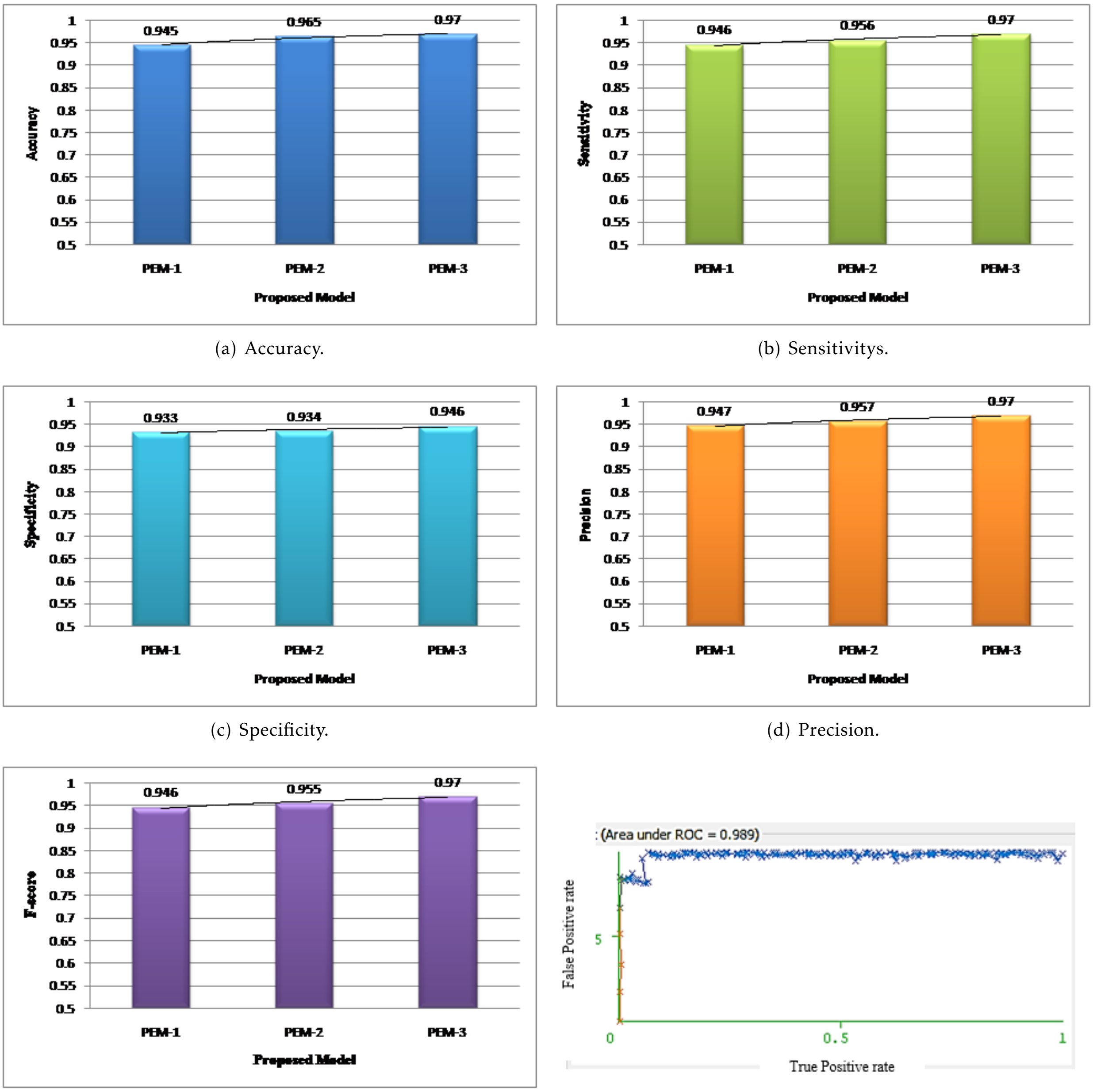
Table 8. Accuracy of proposed MGA-PEM model

Actual Vs Predicted	PEM1	PEM2	PEM3
Accuracy	0.945	0.965	0.970
Sensitivity(TPR)	0.946	0.956	0.970
Specificity(1-FPR)	0.933	0.934	0.946
Precision	0.947	0.957	0.970
F-score	0.946	0.955	0.970
ROC Area	0.940	0.989	0.958

Table 9. Comparison of classification accuracy of proposed and Existing Feature Selection Technique

Feature Selection	No. of Features	PEM-1	PEM-2	PEM-3
PCA	34	68.13%	2.94%	7.84%
CSEBF	78	81.77%	83.25%	82.75%
Proposed MGA	46	94.58%	96.56%	97.04%

In the fifth step of the experience, the classification accuracy of the proposed and existing feature selection techniques was compared. The MGA algorithm was shown to perform better results in comparison to other feature selection techniques.
Table 9 shows the comparison of classification accuracy of the proposed and existing feature selection techniques.
Based on the results obtained in this study, one can conclude that the proposed MGA-PEM presents better results in terms of performance measures. Also, the proposed algorithm is capable to select or reduce the features from the original feature space of a Gene Expression dataset.
The proposed work is compared with similar types of work available in the literature, as illustrated in Table 10. Among these works, the proposed model MGA-PEM is better, producing the highest accuracy with the least number of features (46) and being efficient and robust.
The acronyms used in Table 10 are as follows:
SVM = Support Vector Machine
K-NN = K Nearest Neighbor
RF = Radiofrequency
RBF = Radial Basis Function
GA = Genetic Algorithm
BN = Bayes Network
RBFN = Radial Basis Function Network
DT-J48 = Decision Tree-J48
DT-RF = Decision Tree-Random Forest
PEM = Proposed Ensemble Model.



(a) Accuracy.

(b) Sensitivitys.

(c) Specificity.

(d) Precision.

(e) F-score.

(f) ROC Area.

Figure 5. Various performance measures of proposed model

In our study we have reduced the optimal number of features then we obtained the high classification accuracy as compared to previous work done by different authors.

7. Conclusion

Gene expression data set of lung malignant growth is extremely important in the field of clinical science. Characterization and determination strategies have a crucial role in recognizing precisely a disease. This facilitates the analysis process and correct diagnosis.

In this paper, arrangements models for the characterization of lung disease informational collection were proposed. The characterization methods have shown some improvements as the quantities of highlights or features were diminished.

The proposed group model upheld the Intersection of the current algorithm (GA) based Classification (BN, RBFN, DT-J48 and DT-RF), known as MGA-PEM.

THE MGA-PEM has offered higher grouping exactness contrasted with any or all elements of the lung disease

Table 10. Comparative analysis of the proposed model with the different existing technique

Technique Used	Accuracy	Remarks
SVM, K-NN, and RF	90.70% with RF	Boruta is used for Feature Selection
SVM (RBF), SVM (linear) Ensemble-SVM SVM (linear)-En1	96.2% with SVM (linear	256 features used
GA, BN, RBFN, DT-J48+DT-RF Ensemble Techniques	PEM-1: 94.5% PEM-2: 96.56% PEM-3: 97.04%	46 features used

dataset and existing component characterization procedures.

The grouping calculations Bayes Net (BM) and Radial Basis Function Network (RBFN) are very likely to forecast execution regarding the characterization precision with the proposed MGA-PEM.

The most noteworthy exactness acquired by BN and RBFN classifiers inside the instance of MGA.PEM has 97.04 with 46 highlights.

Alternative robust and computationally economical models are going to develop and MGA-PEM is enforced in the alternative dataset.

The collected gene expression data set is a secondary data set with a low number of instances like 203.

In the future, if the number of samples increases, then the effect of classification accuracy may increase, which would justify our model. We will also develop a new hybrid technique and integrate it into the proposed model to achieve high accuracy with a low false results rate.

References

[1] Song, N., & Wang, K. (2015). Design and Analysis of Ensemble Classifier for Gene Expression Data of Cancer. Journal of Clinical & Medical Genomics, 3(2), 1–7. https://doi.org/10.4172/2472-128x.1000134

[2] Boulesteix, A. L., Strobl, C., Augustin, T., & Daumer, M. (2008). Evaluating microarray-based classifiers: An overview. Cancer Informatics, 6(2007), 77–97. https://doi.org/10.4137/cin.s408

[3] Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association, 97(457), 77–86. https://doi.org/10.1198/016214502753479248

[4] Bonev, B., Escolano, F., & Cazorla, M. (2008). Feature selection, mutual information, and the classification of high-dimensional patterns: Applications to image classification and microarray data analysis. Pattern Analysis and Applications, 11(3–4), 309–319. https://doi.org/10.1007/s10044-008-0107-0

[5] Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. Machine Learning, 37(3), 297–336. https://doi.org/10.1023/A:1007614523901

[6] Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., … Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine, 7(6), 673–679. https://doi.org/10.1038/89044

[7] Sathyadevi, G. (2011). Application of CART algorithm in hepatitis disease diagnosis. International Conference on Recent Trends in Information Technology, ICRTIT 2011, 1283–1287. https://doi.org/10.1109/ICRTIT.2011.5972349

[8] Roslina, A. H., & Noraziah, A. (2010). Prediction of hepatitis prognosis using support vector machines and wrapper method. Proceedings - 2010 7th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2010, 5(Fskd), 2209–2211. https://doi.org/10.1109/FSKD.2010.5569542

[9] Huang, S. H., Wulsin, L. R., Li, H., & Guo, J. (2009). Dimensionality reduction for knowledge discovery in medical claims database: Application to antidepressant medication utilization study. Computer Methods and Programs in Biomedicine, 93(2), 115–123. https://doi.org/10.1016/j.cmpb.2008.08.002

[10] Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., … Robles, V. (2006). Machine learning in bioinformatics. Briefings in Bioinformatics, 7(1), 86–112. https://doi.org/10.1093/bib/bbk007

[11] Inza, I., Larrañaga, P., Blanco, R., & Cerrolaza, A. J. (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. Artificial Intelligence in Medicine, 31(2), 91–103. https://doi.org/10.1016/j.artmed.2004.01.007

[12] https://www.who.int/news-room/fact-sheets/detail/cance

[13] Zhu, S., Wang, D., Yu, K., Li, T., & Gong, Y. (2008). Feature selection for gene expression using model-based entropy. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 7(1), 25-36. DOI: 10.1109/TCBB.2008.35

[14] Rajagopal, S., Kundapur, P. P., & Hareesha, K. S. (2020). A Stacking Ensemble for Network Intrusion Detection Using Heterogeneous Datasets. Security and Communication Networks, 2020. https://doi.org/10.1155/2020/4586875

[15] Ragunthar, T., & Selvakumar, S. (2019). Classification of Gene Expression Data with Optimized Feature Selection, (2), 4763–4769. https://doi.org/10.35940/ijrte.B1845.078219

[16] Ramroach, S., John, M., & Joshi, A. (2019). The Efficacy of Various Machine Learning Models for Multi-class Classification of RNA-Seq Expression Data. Advances in Intelligent Systems and Computing, 997, 918–928. https://doi.org/10.1007/978-3-030-22871-2_65

[17] Sahu, S. K., & Shrivas, A. K. (2018). Analysis and Comparison of Clustering Techniques for Chronic Kidney

Disease With Genetic Algorithm. International Journal of Computer Vision and Image Processing, 8(4), 16–25. https://doi.org/10.4018/IJCVIP.2018100102

[18] Zahoor, J., & Zafar, K. (2020). Classification of Microarray Gene Expression Data Using an Infiltration Tactics Optimization (ITO) Algorithm. Genes, 11(7), 819. doi:10.3390/genes11070819

[19] Chen, Q., Meng, Z., & Su, R. (2020). WERFE: A Gene Selection Algorithm Based on Recursive Feature Elimination and Ensemble Strategy. Frontiers in Bioengineering and Biotechnology, 8. doi:10.3389/fbioe.2020.00496

[20] Bharati, S., Rahman, M. A., & Podder, P. (2018, September). Breast cancer prediction applying different classification algorithm with comparative analysis using WEKA. In 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT) (pp. 581-584). IEEE.

[21] Parimala, R., & Nallaswamy, R. (2011). A Study of Spam E-mail classification using Feature Selection package. Global Journal of Computer Science and Technology, 11(7), 45–54.

[22] Ashraf, M., Chetty, G., & Tran, D. (2013). Feature selection techniques on thyroid, hepatitis, and breast cancer datasets. International Journal on Data Mining and Intelligent Information Technology Applications, 3(1), 1. DOI: 10.4156/IJMIA.vol3.issue1.1

[23] Hall, M. (1999). Correlation-based Feature Selection for Machine Learning. Methodology, 21i195-i20(April), 1–5.

[24] Holland, J.H. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence; MIT Press: Cambridge, MA, USA, 1975

[25] Alabed A., Kambhampati C., Gordon N. (2020) Genetic Algorithms as a Feature Selection Tool in Heart Failure Disease. In: Arai K., Kapoor S., Bhatia R. (eds) Intelligent Computing. SAI 2020. Advances in Intelligent Systems and Computing, vol 1229. Springer, Cham. https://doi.org/10.1007/978-3-030-52246-9_38

[26] Bhattacharjee, a, Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., … Meyerson, M. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A, 98(24), 13790-5. https://doi.org/10.1073/pnas.191502998

[27] Nagalla, R., Pothuganti, P., & Pawar, D. S. (2017). Analyzing Gap Acceptance Behavior at Unsignalized Intersections Using Support Vector Machines, Decision Tree and Random Forests. Procedia Computer Science, 109(2016), 474–481. https://doi.org/10.1016/j.procs.2017.05.312

[28] Ross, J., Morgan, Q., & Publishers, K. (1994). Book Review: C4.5: Programs for Machine Learning, 240, 235–240.

[29] Velichkov, B., Koychev, I., & Boytcheva, S. (2019). Deep learning contextual models for prediction of sport events outcome from sportsmen interviews. International Conference Recent Advances in Natural Language Processing, RANLP, 2019-Septe, 1240–1246. https://doi.org/10.26615/978-954-452-056-4_142

[30] Pearl, J. (1985, August). Bayesian netwcrks: A model cf self-activated memory for evidential reasoning. In Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine, CA, USA (pp. 15-17).

[31] Bouzembrak, Y., & Marvin, H. J. P. (2018). Impact of drivers of change, including climatic factors, on the occurrence of chemical food safety hazards in fruits and vegetables: A Bayesian Network approach. Food Control, 97(August 2018), 67–76. https://doi.org/10.1016/j.foodcont.2018.10.021

[32] Cheng, J., Greiner, R., Kelly, J., Bell, D., & Liu, W. (2002). Learning Bayesian networks from data: An information-theory based approach. Artificial Intelligence, 137(1–2), 43–90. https://doi.org/10.1016/S0004-3702(02)00191-1

[33] Zadeh, M., & Mahmoudi, J. (2017). Discrimination between Earthquakes and Explosions Using Artifi cial Neural Networks. American Journal of Biometrics and Biostatistics, 2(1), 1–6. https://doi.org/10.1016/j.polymer.2015.09.054

[34] Vera Candioti, L., De Zan, M. M., Cámara, M. S., & Goicoechea, H. C. (2014). Experimental design and multiple response optimization. Using the desirability function in analytical methods development. Talanta, 124, 123–138. https://doi.org/10.1016/j.talanta.2014.01.034

[35] Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. Clinical Chemistry, 39(4), 561–577. https://doi.org/10.1093/clinchem/39.4.561