International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v10i8.5664
***Article Received**: 10 June 2022 **Revised**: 15 July 2022 **Accepted**: 30 July 2022 **Publication**: 31 August 2022*

EapGAFS:Microarray Dataset for Ensemble Classification for Diseases Prediction

Peddarapu Rama Krishna[1,*], Dr. Pothuraju Rajarajeswari[2]
[1]Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India.
[2]Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India.
Email:peddarapuramakrishna@gmail.com

Abstract

Microarray data stores the measured expression levels of thousands of genes simultaneously which helps the researchers to get insight into the biological and prognostic information. Cancer is a deadly disease that develops over time and involves the uncontrolled division of body cells. In cancer, many genes are responsible for cell growth and division. But different kinds of cancer are caused by a different set of genes. So to be able to better understand, diagnose and treat cancer, it is essential to know which of the genes in the cancer cells are working abnormally. The advances in data mining, machine learning, soft computing, and pattern recognition have addressed the challenges posed by the researchers to develop computationally effective models to identify the new class of disease and develop diagnostic or therapeutic targets. This paper proposed an Ensemble Aprior Gentic Algorithm Feature Selection (EapGAFS) for microarray dataset classification. The proposed algorithm comprises of the genetic algorithm implemented with aprior learning for the microarray attributes classification. The proposed EapGAFS uses the rule set mining in the genetic algorithm for the microarray dataset processing. Through framed rule set the proposed model extract the attribute features in the dataset. Finally, with the ensemble classifier model the microarray dataset were classified for the processing. The performance of the proposed EapGAFS is conventional classifiers for the collected microarray dataset of the breast cancer, Hepatities, diabeties, and bupa. The comparative analysis of the proposed EapGAFS with the conventional classifier expressed that the proposed EapGAFS exhibits improved performance in the microarray dataset classification. The performance of the proposed EapGAFS is improved ~4 – 6% than the conventional classifiers such as Adaboost and ensemble.

Keywords: Microarray datasets, ensemble classifier, genetic algorithm, diseases, classification.

1. Introduction

In recent years, the epidemic in breast cancer, diabetes, liver disorder, prostate cancer, colon tumor, obesity and many other heart diseases has become a challenge to global health. The dreadful diseases like cancer [1] often proves to be life-altering, life threatening and fatal. Most often their symptoms stem from a genetic basis [2] and a host of challenges demand for the prevention, diagnosis, treatment and cure of these diseases. As medicine plays a great role in saving human life, medical data classification has remained as one of the. Medical data [3,4] straddles on clinical data and genomic data, which is the cornerstone of biomedical informatics. The domain of biomedical informatics has emerged due to the cross fertilization of bioinformatics, medical informatics and clinical genomics [5]. As bioinformatics is emerging technology associated with the processing of the issues associated with the data in terms of collection, analysis, processing and retrieval based on evaluation of function and structure in the biological system. Biomedical informatics comprises of different application for biology, healthcare technology, applied mathematics to exhibits significant solution for the clinical data. Because of their common originality, they mutually influence each other on many occasions. As biomedical database becomes more and more voluminous, there is urgent need to develop, create and apply new algorithms to model, manage and interpret this information.

The application of data mining, machine learning and evolutionary computing techniques to the field of biomedical informatics can handle large amount of biological data and extract the hidden knowledge from it. The computational complexity is greatly reduced by adopting gene selection methods [6]. Various types of evolutionary computing techniques are often employed for gene selection. Machine learning is a consortium of artificial intelligence and computational intelligence. Their role model is the human mind based on which they target to develop intelligent machines to solve real life problems. It comprises neuro

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v10i8.5664
***Article Received**: 10 June 2022 **Revised**: 15 July 2022 **Accepted**: 30 July 2022 **Publication**: 31 August 2022*

computing, fuzzy logic, genetic computing, probabilistic reasoning including genetic algorithm, belief networks, learning theory which collectively provide a foundation for the conception, design and deployment of intelligent tools [7]. The general problems in the domain of biomedical informatics is to classify the medical data and genomic data. Small medical data classification aims to identify and diagnosis the diseases in the patient. Additionally, through dimensionality reduction classification is performed for the genomic data with reduction of the feature set evaluated with the feature extraction or selection process to perform the classification process.

Machine learning strongly relies on evolutionary computing techniques for determining the parameters those are processed with the neural network, neurofuzzy hybrid network and genome data for appropriate selection of genes. Through implementation of the hybrid neural system different machine learning techniques are evolved such as evolutionary neural network, hybrid neural fuzzy system, neuro fuzzy system, extreme learning machine and so on [8]. Those machine learning model exhibits the significant performance in the real time environment for the complex applications. The performance is achieve and improved through imprecision, approximation, partial truth and uncertainty [9]. Though biological data is often associated with noisy and missing samples, the intelligent techniques handle them efficiently. In many cases the results are found to be promising when various medical data, micro array medical data are classified by these models [10,11].

DNA Microarray is the tool, which helps the researcher to explore the expression values of thousands of genes in a single experiment swiftly and competently [12]. There are various types of cells in the human body and each cell consists of copies of the same set of 20,000 genes, but the cell types differ from each other based on the activation of certain genes [13]. By comparing the gene expression profiles of any cell type, one could learn what makes the two cells different from each other [14]. Using DNA microarray, researchers can point out the difference between two different cell types in a single experiment. A usual microarray experiment uses the hybridization of a mRNA molecule to the DNA template from which it was originated. An array is constructed using many DNA samples. The expression levels of various genes were measured by the amount of mRNA bound to each site on the array. This number may run in thousands. With the collection of all the data, a gene expression profile for the cell is generated [15].

With the development of gene expression datasets, the opportunity to develop different computational models for disease diagnosis, prognosis and prediction of a disease is widening. The advances in data mining, machine learning, soft computing and pattern recognition have addressed the challenges posed by the researchers to develop computationally effective models to identify the new class of disease and developing diagnostic or therapeutic targets [16]. The major motivation of this research is to extract relevant information from this high dimensional data to have clinical decision support. There are several different data analysis difficulties which have to be overcome to extract knowledge from this high dimensional data which can be served to be a clinical application. Cancer is a principal cause of death in all over the world. In 2012 there were 14 million new cases and 8.2 million cancer-related deaths worldwide according to the cancer statistics found by national cancer institute (NCI). Based on the data collected in the year 2010-2012 approximately 39.6% of people will be diagnosed with cancer at some points during their lifetimes. Therefore, a cautious analysis of Microarray data can help early diagnosis and prognosis of cancer which can be a significant contribution to the society [17].

1.1 Contribution and Organization of the Work

This paper presented a microarray dataset processing algorithm for the disease's prediction in the dataset. The proposed technique is termed as EapGAFS model for the prediction model. The specific contribution of the proposed EapGAFS model is presented below:

- The proposed EapGAFS uses the rule set mining for the estimation of the attributes features in the microarray dataset.
- The extracted features with the rule set are processed with the genetic algorithm model for the optimization of the features in the data for improving accuracy in the data classification and prediction.
- With the developed EapGAFS model extracted features ensemble classifier is applied for the processing and classification of the disease's prediction microarray datasets.
- The analysis is based on the consideration of the four diseases datasets such as breast cancer, Hepatities, diabeties, and bupa. Through the extensive analysis the proposed EapGAFS model performance is evaluated for the different disease's datasets.
- The comparative analysis stated that the proposed EapGAFS was evaluated with conventional adaboost and ensemble classifier. The proposed EapGAFS exhibits improved performance in the microarray dataset classification. The performance of the proposed EapGAFS is improved ~4 – 6% than the conventional classifiers such as Adaboost and

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v10i8.5664
***Article Received**: 10 June 2022 **Revised**: 15 July 2022 **Accepted**: 30 July 2022 **Publication**: 31 August 2022*

ensemble.

This paper is organized as follows: Section 2 provides the related works associated the with the disease prediction microarray datasets. The proposed EapGAFS model for the disease prediction and classification is presented in section 3. In section 4 the results obtained for the proposed EapGAFS is presented followed by the overall conclusion is presented in section 5.

2. Related Works

The complete project of type makes a speciality of the proper identity of those features which clearly make a contribution towards growing the perfect classifiers. Naturally the least sizable or faulty functions get discarded and this manner is called as size reduction. The trouble of excessive dimensionality is regularly tackled through person distinctive subspaces of hobby. However, consumer identity of the subspaces is errorprone within the absence of prior domain know-how. Another manner to cope with curse of dimensionality is to apply a dimensionality discount approach to the dataset. Microarray technology produces massive datasets with gene expression values for heaps of genes (6000-60000) in a cell combination [18]. Hence, it turns into economically prohibitive to have a huge pattern length. This phenomenon is known as as a curse of dimensionality in which samples (n) << the variety of features (p). To conquer this trouble, microarray clinical datasets need measurement discount [19]. Dimensionality discount methods are broadly categorized into types i.E. Feature extraction and function selection [20]. Feature extraction algorithms are the strategies or strategies that remodel the original characteristic set. In different phrases, given a nXd pattern matrix A (n factors in a d-dimensional space), a nXm sample matrix B is being derived, such that m << d where B=AH and H is a d X m transformation matrix.

Feature selection is classified as semi-supervised, supervised and unsupervised approach for prediction of the relay in the class [21]. The feature selection approach comprises of the wrapper, filter and embedded model for processing. The filter model exhibits the intrinsic performance analysis through the statistics. In those technique, all features are scored and ranked based on the statistical criteria. Based on the scoring the ranking is performed for the function values with appropriate scoring values. Through strategic choices the predominat analysis is performed with elimination of the classifier interaction, elimination of the functional dependencies.

The classical filter techniques are correlation characteristic selection, fast correlation based totally clear out, records advantage emedy, minimal Redundancy Maximum Relevance (mRMR), aid vector gadget recursive function removal (SVMRFE), sequential ahead choice (SFS) and sequential backward elimination (SBE) and so forth. Amongst all the strategies, SFS and SBE are drastically used due to their simplicity and low computational overhead. But additionally, they have their own barriers. The fundamental disadvantage of the sequential search technique is the nesting impact i.E. In backward search while a function is deleted it can't be reselected and in ahead search when a characteristic is chosen, it cannot be deleted [22]. The wrapper version [23] makes use of a predictive accuracy of a special getting to know algorithm to determine the fine of selected functions. This approach is computationally pricey to run massive datasets having huge wide variety of capabilities. The embedded version bridges the distance among these fashions with the aid of taking the blessings from both the techniques [24]. The modern trend of function selection has additionally adopted many hybrids or ensemble strategies [25] that have proven superior type performance and strong characteristic choice outcomes. Finally feature selection method facilitates in selecting a most useful subset of relatively discriminating functions from the original function set with none transformation. Hence, function selection is superior to characteristic extraction in terms of better clarity and interpretability. Dimensionality reduction helps in the class of microarray clinical datasets by enhancing its accuracy.

In order to triumph over the restrictions of the traditional strategies, the stochastic search strategy is followed in which a few randomness is added in the seek process and the characteristic selection manner will become much less touchy to the particular dataset. Literature survey well-known shows that numerous competing techniques are hired for the challenge of feature selection, whose last purpose is to understand greater compact and better exceptional function subsets. In this process the significance of the functions are evaluated, their consistencies are measured and a pre-determined seek algorithm is used together with a classifier. The first-rate of a function subset is measured in phrases of category accuracy. The characteristic subset contributing maximum classification accuracy with minimal quantity of functions is considered to be maximum most appropriate. Various sorts of evolutionary algorithms (EAs) are used for feature selection. These stochastic seek methods mimic the metaphor of natural organic evolution and function on capacity answers as a consequence of the procedure of natural genetics. The first-class solution is received on the stop. Many researchers from the Artificial Intelligence network have incorporated EAs and ANNs. EAs are carried out to microarray class to observe the most efficient, or near top of the line, subset of predictive genes on the complex and big areas of feasible gene units. The most

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v10i8.5664
***Article Received**: 10 June 2022 **Revised**: 15 July 2022 **Accepted**: 30 July 2022 **Publication**: 31 August 2022*

famous stochastic methods of characteristic selection are genetic algorithm, simulated annealing, ant colony optimization, particle swarm optimization, differential evolution, bacterial foraging optimization, harmony search, cuckoo search, firefly, bat algorithm and cat swarm optimization [26].

When educated with gradient descent learning strategies, neural community-based fashions be afflicted by gradual gaining knowledge of, clean convergence to nearby minima and overfitting. To keep away from these obstacles, Huang et. Al. Proposed extreme gaining knowledge of device (ELM) which proved its superiority in phrases of extremely rapid studying velocity and good generalization potential. Literature survey exhibits that ELM has originated from RVFL [27].

3.Main EaspGAFS Extreme Ensemble Classifier Model

An evolutionary algorithm is considered a major point in health care applications. It offers the decision based on the data received from the patient's body and other information. In other words, evolutionary model is stated as analyze of microarray data that are generated electronically for the aid of the decision-making process. Nowadays, a vast range of applications uses an evolutionary algorithm for the processing of medical data of patients for the provision of desired medical information. The evolutionary system has been classified into two categories such as knowledge-based and non-knowledge-based. To extensive uses, evolutionary algorithm the improvement in AI or ML is performed for drastic improvement in the process. The main aim of the evolutionary based ensemble model system is the provision of patient data accurately for deciding on health care. Also, the incorporation of evolutionary model minimizes the cost with the improved efficiency and reduced inconvenience of patients. The general block for evolutionary process in medical data is presented in figure 1.

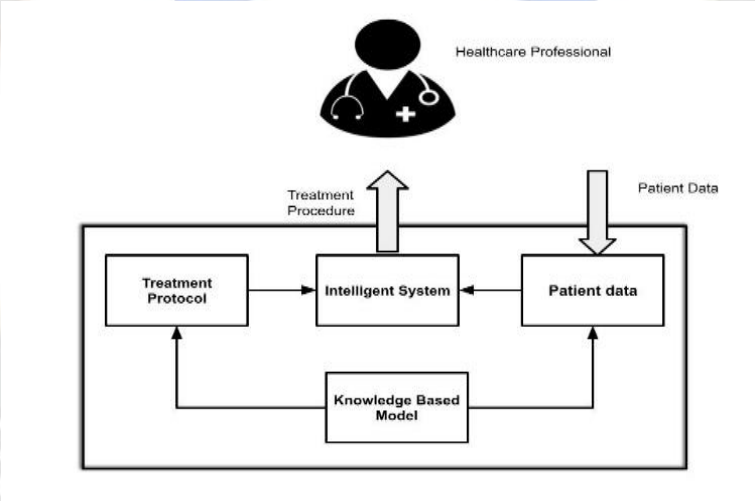


Fig.1. Evolutionary Model

The implementation of ensemble and genetic algorithm within evolutionary algorithm offers effective information quickly and cost-effective manner. Also, it provides an adequate decision-making process in a safe, effective, and acceptable manner. This paper concentrated on the construction of an effective ensemble based evolutionary algorithm for the microarray dataset for disease prediction. As the proposed EapGAFS incorporates the basic structure of the EapGAFS with consideration of four layers with the integration of parallel processing of neurons. The process involved in AI is illustrated in figure 2.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v10i8.5664
***Article Received**: 10 June 2022 **Revised**: 15 July 2022 **Accepted**: 30 July 2022 **Publication**: 31 August 2022*

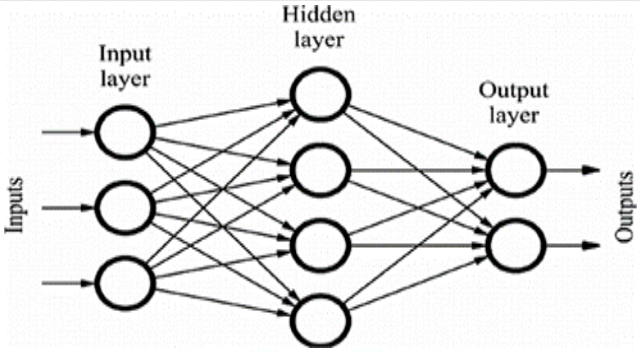


Fig.2. Ensemble Model with evolutionary process

Initially, upon successive layers, each layer received information before transferring to the next layer. Finally, the output is obtained at the last layer of the network. The proposed EapGAFS involved is an automated diagnosis of microarray dataset information of patients. The proposed method incorporates an AI-based smart technique for information gathering from microarray datasets. The incorporated microarray dataset is involved in the transmission of information about the disease diagnosis. The incorporated Genetic algorithm use various attributes in the microarray dataset to record the information about disease. The information acquired from the microarray dataset is extracted and processed with the ensemble evolutionary model.

This paper developed a EapGAFS architecture integrated with Genetic algorithm for the diagnosis of disease in the microarray datasets. The proposed architecture incorporates an evolutionary algorithm integrated with the ensemble classifier for processing collected medical information. The proposed model concentrates on the real-time processing of data to provide adequate information and knowledge to the with ensemble machine learning model. The proposed model uses GA algorithm for attribute clustering and processed information are evaluated with the ensemble classifier model. The GA model uses a ensemble classifier for the classification of the microarray dataset. The overall process in the proposed EapGAFS for the microarray dataset are presented in figure 3.

1. Genetic algorithm for data clustering
2. Aprior model for attribute evaluation with rule mining
3. Ensemble classifier
4. Classification model

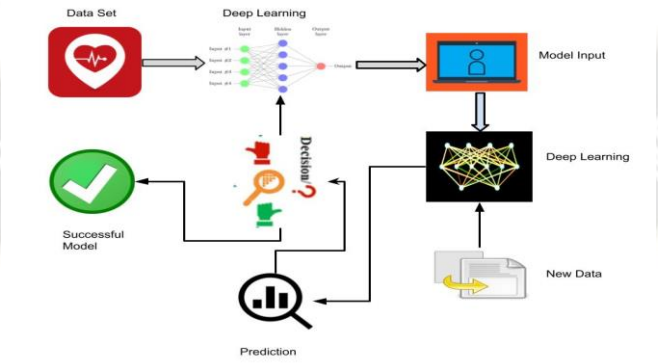


Fig.3. Overall Process of EapGAFS

The proposed EapGAFS comprises of the module such as a genetic algorithm, aprior learning and ensemble classifier. As, the proposed EapGAFS uses the GA and ensemble classifier it effectively performs the data clustering and analysis. The overall process in EapGAFS is presented in graphical form in figure 4.

3.1 Genetic Algorithm (GA) based Feature Selection in Microarray Datasets

The GA uses the Chromosome Representation for binary encoding technique has been used for GA-based feature selection. The string represents the presence or absence of attributes in the dataset. Fitness Function: The formulation of the genetic algorithm is to maximize the classification accuracy and minimize the number of features selected. Here, ensemble is used as a classifier to measure the fitness of each chromosome. The fitness is calculated based on equation (1).

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v10i8.5664
Article Received: 10 June 2022 Revised: 15 July 2022 Accepted: 30 July 2022 Publication: 31 August 2022

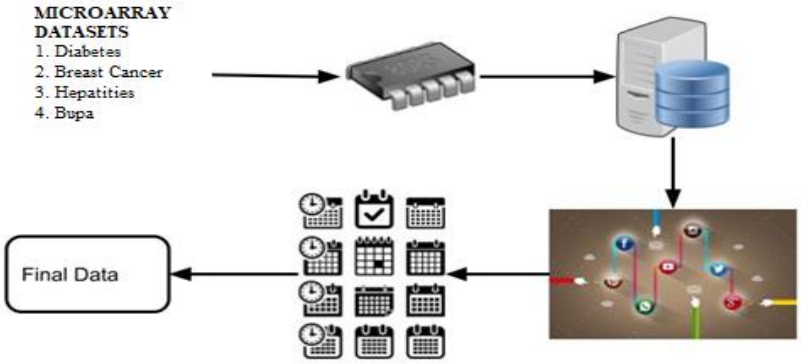


Fig.4. Block Diagram of EapGAFS

$$F = p * Accuracy(f) + (1 - P) * \frac{1}{1+size(f)} \quad (1)$$

In above equation (1) the feature vector is denoted as f for the selected microarray dataset, the parameter is denoted as P between the range [0, 1]. The classification accuracy is evaluated based on two objectives to control the feature subset. The measured accuracy value is computed as in equation (2)

$$Accuracy = \frac{s}{c} * 100 \quad (2)$$

In the proposed EapGAFS the correctly classified sample are represented as s and total test sample data is denoted as c. With the proposed EapGAFS comprises of the set of selection, crossover, mutation and evaluation. The steps adopted in the EapGAFS for the implementation of the GA in feature selection is presented in steps 1.

Step 1: Steps in GA for feature selection

Input: Train the dataset and attributes are represented as p from the feature pool

Output: Compute the subset of optimal features

Step 1: Generate the initial population randomly from the set of N chromosomes.

Step 2: Formulate the population sequence from the estimated population using the following steps

Step 2.1: Compute the microarray population fitness population using the equation (2)

Step 2.2: Evaluate the fitness for the members

Step 2.3: Consider the present population as elite for the generation of the population

Step 2.4: Perform the crossover operation from the parent from the generated offspring

Step 2.5: Compute probability of mutation for the every offspring in every locus.

Step 2.6: Replace the present population with children in the next iteration

Step 3: Stop in the stopping criteria.

Step 4: Best individual value need to be return

Step 5: Compute the feature subset optimal values.

3.2 Group Genetic Algorithm (GGA) for Microarray Clustering

The developed group genetic algorithm (GGA) uses the attribute clustering in the microarray dataset. Compute the subset into K features with the predefined constant value. Consider the feature set T for features n is represented as {g1, g2,…,gn}. With the defined set random population are generated for the K features those are generated form nonempty groups are represented in equation (3)

$$G = G_1, G_2 \dots \dots, G_k \quad (3)$$

The assigned features N are assigned randomly for each group. The assigned features for every empty group are selected random manner from the set those splited randomly in two groups denoted as $\left\{G_l \| G_l| > \frac{N}{K}, G_l \in G\right\}$. The classification accuracy for the subsets attributes are evaluated based on the training dataset. The accuracy for the proposed EapGAFS is defined as in equation (4)

$$Acc(A_i) = \sum_{p=1}^{NA} \frac{sub\ Acc(T_p)}{NA} \quad (4)$$

In above equation (4), attribute combination is denoted as NA

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v10i8.5664
***Article Received**: 10 June 2022 **Revised**: 15 July 2022 **Accepted**: 30 July 2022 **Publication**: 31 August 2022*

and T_p is denoted as the combination of the possible attributes. The balance cluster in the group A_i is represented as in equation (5)

$$balance(A_i) = \sum_{i=1}^{k} -\frac{|grp(i)|}{N} log \frac{|grp(i)|}{N} \quad (5)$$

The attribute number in i[th] group is represented as N features for the computation of the fitness function as denoted in equation (6)

$$f(A_i) = Acc(A_i) * [balance(A_i)]^a \quad (6)$$

Here, the cluster accuracy relative value is balanced with the proposed EapGAFS for cluster balance in the microarray dataset. The steps involved in the for clustering are

1. Select the base chromosome form the random selected chromosomes (C1).
2. Insert another base chromosome (C2) in the groups
3. Eliminate the formed new chromosomes (Cnew) in duplicate value
4. If $|C_{new}| < K$ then, generate the randomly selected group as $\left\{G_l \| G_l| > \frac{N}{K}, G_l \in G\right\}$ those splitted in to two groups.
5. Else if $|C_{new}| > K$ perform the wheel selection strategy for each group those group has better removal probability value.

There are a number of kernels used in SVM classifier such as linear, polynomial, RBF and sigmoid. In this work, C-SVM with a linear kernel is used. In C- SVM, training involves the minimization of the error function is evaluated using equation (7)

$$\frac{1}{2} w^T w + C \sum_{i=1}^{N} \varepsilon(i) \quad (7)$$

Subject to the constraints represented in equation (8):

$$y(i)\left(w^T \emptyset\left(x(i)\right) + b\right) \geq 1 - \varepsilon(i) \quad (8)$$

And $\varepsilon(i) \geq 0, i = 1, \ldots\ldots, N$ with the capacity constant denoted as C with vector coefficient value of w for the constant b and non-separated data is denoted as $\varepsilon(i)$. THe training sample N index sample is computed using rule mining with the transaction subset item value is denoted as $T = \{t_1, t_2, t_3 \ldots\ldots t_n\}$. The support rule for the EapGAFS rule is computed as in equation (9) and equation (10)

$$\boldsymbol{Support}\ (\boldsymbol{X \rightarrow Y}) = \frac{\boldsymbol{Support\_Count(X \cup Y)}}{|\boldsymbol{T}|} \quad (9)$$

$$\boldsymbol{Confidence}\ (\boldsymbol{X \rightarrow Y}) = \frac{\boldsymbol{Support(X \cup Y)}}{\boldsymbol{Support}\ (\boldsymbol{X})} \quad (10)$$

The computed ration for the support between independent characteristics of the X and Y denoted as in equation (11)

$$\boldsymbol{lift}\ (\boldsymbol{X \rightarrow Y}) = \frac{\boldsymbol{Support(X \cup Y)}}{\boldsymbol{Support}\ (\boldsymbol{X}) * \boldsymbol{Support}\ (\boldsymbol{Y})} \quad (11)$$

The proposed EapGAFS rule estimate the confidence satisfaction based on the minimal threshold confidence value and lift >=1 to generate strong relationship. The overall architecture in proposed EapGAFS is presented as follows.

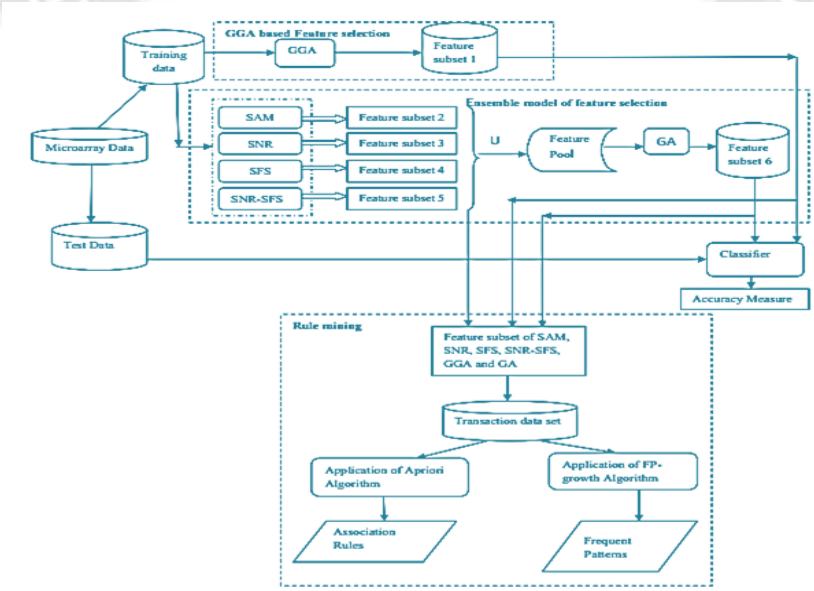


Fig.5. Overall Architecture of the EapGAFS

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v10i8.5664
Article Received: 10 June 2022 Revised: 15 July 2022 Accepted: 30 July 2022 Publication: 31 August 2022

The generated initial population size is denoted as t with the learners as (t= 1, 2, 3 ... n) with the designed variable as S. for the subjects (s=1, 2, 3…..m). The learners each value in the subject is evaluated with the objective function represented as in equation (12)

$$\text{Minimum f(x)} = \sum_{i=1}^{n}[x_i^2 - 10\cos(2\pi x_i) + 10] \quad (12)$$

In the learning phase the best learner is computed those involved in improvement in the learners mean. For every subjects j=1, 2, 3, …….m the ith iteration is evaluated based on the mean iteration is computed as in equation (13) - (15)

$$\boldsymbol{Difference = r(M_{new}, s - T_F M_s)} \quad (13)$$

$$\boldsymbol{D_{Chebyshev}(x_i, x_j) = max(|x_i - x_j|)} \quad (14)$$

$$\boldsymbol{X_{new} = f(x) + D_{Chebyshev}(x_i, x_j)} \quad (15)$$

Through the obtained value the present value is updated based on the current tvalue with objective function X'_{new}. The formulated new objective function is computed for the function value. In the training phase the interaction with the other learners are estimated. The training process is presented as follows:

Step 1: Select another learner randomly x_j, such that $i \neq j$

Step 2: If the objective function of the new learner is greater than the old learner

i.e $f(x_j) < f(x_j)$ then go to step 2 else go to step 4

Step 3: Calculate the mean of the new learner as $X''_{new,i} = X''_{new,i} + r_i(X_i - X_j)$

Step 4: Update the mean value of the learners as $X''_{new,i} = X''_{new,i} + r_i(X_j - X_i)$

Step 5: Accept $X''_{new,i}$ if function value is better than the previous value

The Evolutionary structure is involved in the determination diseases for with the microarray dataset. The model with ensemble modle compute the GA for the clustering of the features in the microarray dataset. Initially, the ensemble model incorporates hidden later or feed-forward to machine learning model. The overall architecture of the proposed EapGAFS is presented in figure 6.

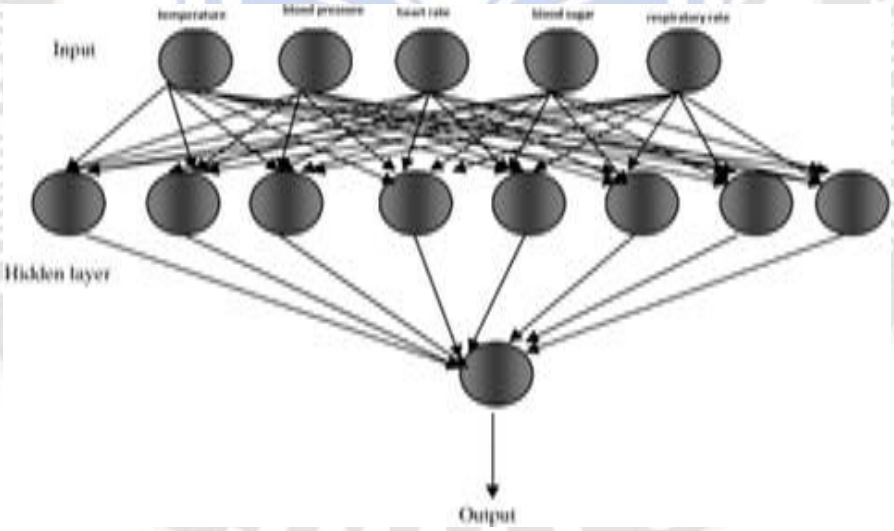


Fig.6.Architecture of Ensemble Classifier

The hidden layer comprises 8 layers in which the last layer of network output layers comprises the node. Also, the EapGAFS exhibits significant developmental applications to evaluate the efficiency of the microarray dataset models. In the next stage, double hidden AI layers are defined to achieve output for the first input layer those vary in accordance to double hidden layer with eight neurons in which the second layer comprises of three layers. As stated, the proposed EapGAFS comprises five layers of medical sensors and the output of microarray dataset. Based on the generated rule hidden neuron layers are processed with the determination of the appropriate neuron size of the hidden layer in ensemble model. The neuron selected for the hidden layer are varying in size twice than the input layer of the network. Upon the formulation of the feed-forward network in AI, the neurons activation function is performed in the microarray datset. The proposed EapGAFS uses the activation function sigmoidal, Tanh and hybrid AI.

3.3 Ensemble model for the classification

The proposed EapGAFS uses a back-propagation algorithm for the training and learning of AI. The updated weights of the data collected from the network are stated as $(w_1, w_2, w_3, \ldots\ldots)$. The overall inputs applied over the deep

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v10i8.5664
***Article Received**: 10 June 2022 **Revised**: 15 July 2022 **Accepted**: 30 July 2022 **Publication**: 31 August 2022*

learning network DSS_HeAI is stated in equation (16) as follows:

$$y = sig[\sum_1^N x_N w_N - \partial_N] \quad (16)$$

In the above equation, input is represented as x_n, the weights are denoted as w, and the threshold is computed as ∂_N.

With the application of the sigmoidal activation function, the network is stated in equation (17) – (18)

$$sig(x) = \frac{1}{1+e^x} \quad (17)$$

$$sig(x) = \frac{1}{1+exp[\sum_1^N x_N w_N - \partial_N]} \quad (18)$$

The derivation achieved through the sigmoidal function is denoted in equation (19)

$$\frac{\partial}{\partial x}(sig) = [1 - sig(x)].[sig(x)] \quad (19)$$

The error function estimated for the proposed EapGAFS is presented in equation (20) with computation of total error computation as presented below.

$$TE_{total} = \sum \frac{1}{2}(desire\ (target) - actual)^2 \quad (20)$$

The developed EapGAFS uses backpropagation with an ensemble model for the computation of error in the complete network. The gradient descent method is presented in equation (21) - (23)

$$\partial_L = \frac{\partial}{\partial_{yN}} sig * e_k \quad (21)$$

$$\partial_L = Q[sig\ (y_k)].[1 - sig\ (y_k)].e_k \quad (22)$$

$$TE_0 = \frac{1}{N}\sum Q\left(A_N^{expected} - A_N^{actual}\right)^2 \quad (23)$$

The deep learning gradient for error is represented as in equation (24)

$$\partial_N^{`} = \frac{\partial}{\partial yj}(sig).\left(w_{jk}.\partial_N\right) \quad (24)$$

The change in weights of the network is defined as

$$w_{N+1} = w_N + \Delta w_N \quad (25)$$

$$\Delta w_N = Q.\emptyset.x_N.e \quad (26)$$

Algorithm: Proposed EapGAFS

Initialize the weights of the classifier for input x_t, t = 1,2, …., T
for t = 1 to T;
estimate the microarray dataset as $h_\theta(x)$
compute
$y = \omega^T x$
Estimate sigmoidal analysis for collected data $g\left(\frac{1}{1+e^{-\theta T_x}}\right)$
Compute the probability of variable in the sigmoidal for input variables of P for computation of weak classifier
Compute probability in the network using equation (9) – equation (11)
Estimate the range of probability in the network
$\hat{l}(\theta; x) = \frac{1}{n}\sum_{i=1}^{n} \ln f(x_i|\theta)$
for
$P < 1$ compute as a data diagnosis
else
compute as abnormal.
end for
Compute the ensemble classifier with
$\Delta w_N = Q.\emptyset.x_N.e$
Estimate the classified data in the network model

4. Results and Discussion

Performance Metrics

Accuracy: To evaluate the correctly classified instances for total predicted values to the number of predictors. It is defined in equation (27)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (27)$$

Recall or Sensitivity: Recall provides a ratio of the value of correct prediction to total prediction values. It is defined in equation (28)

$$\text{Recall} = \frac{TP}{TP+FN} \quad (28)$$

Precision: It provides a ratio of true positive values to total predicted values. It is stated in equation (29)

$$\text{Recall} = \frac{TP}{TP+FP} \quad (29)$$

F1 - Score: It is involved in the computation of the ratio between the average mean value of precision and Recall. F1-Score is stated in equation (30)

$$F1_{Score} = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (30)$$

Confusion Matrix: It provides the performance of the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v10i8.5664
***Article Received**: 10 June 2022 **Revised**: 15 July 2022 **Accepted**: 30 July 2022 **Publication**: 31 August 2022*

proposed model with a comparative analysis of actual and predicted values. The analysis is based on the estimation of TP, FN, FP, and TN. It is represented in equation (31)

$$ConfusionMatrix = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \quad (31)$$

Where, True Positive (TP) - is stated as forecast value which is anticipated as positive.

False Positive (FP) - predicted value is considered as negative and converted to positive.

True Negative (TN) - predict as negative and later considered as unpredictable.

False Negative (FN) - predicted as positive and changes to negative.

The dataset considered for the analysis of the proposed EapGAFS are presented in table 1 as follows:

Table 1 Distribution of the datasets

Dataset	Training Samples	Testing Samples	Features	Class
Breast Cancer	498	197	9	2
Bupa	543	143	6	2
Hepatities	78	68	19	2
Diabetes	569	189	8	2

Performance of all the classifiers discussed in this study are evaluated by different measures like training accuracy, testing accuracy, confusion matrix, receiver operating characteristic curve (ROC), sensitivity, specificity, Gmean, and F-score which are explained

4.1 Breast Cancer Dataset

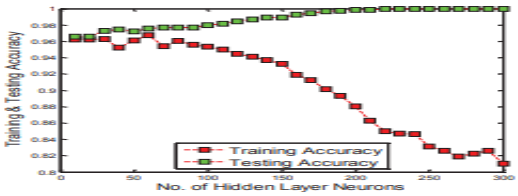


Fig.7. Training and testing accuracy of ELM

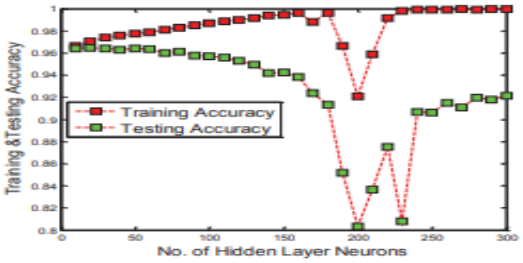


Fig.8.Training and testing accuracy of Adaboost

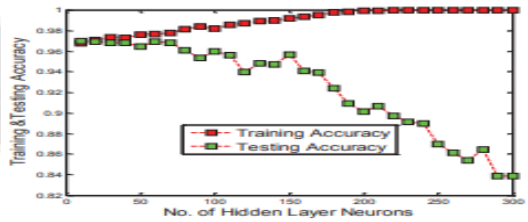


Fig.9. Training and testing accuracy of proposed EapGAFS

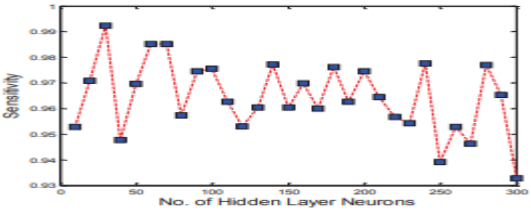


Fig.10.Sensitivity of ELM

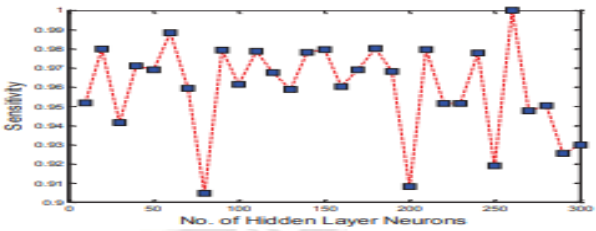


Fig.11.Sensitivity of AdaBoost

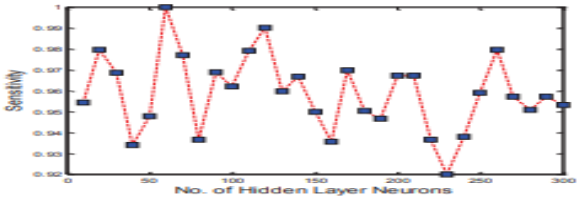


Fig.12.Sensitivity of EapGAFS

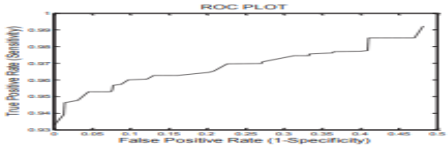


Fig.13. ROC of ELM

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v10i8.5664
Article Received: 10 June 2022 Revised: 15 July 2022 Accepted: 30 July 2022 Publication: 31 August 2022

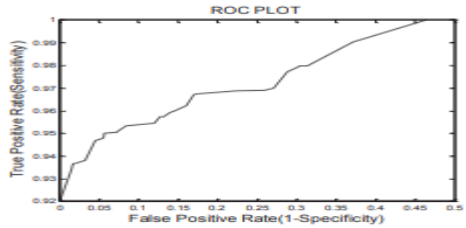


Fig.14. ROC of AdaBoost

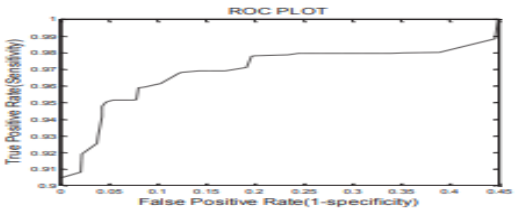


Fig.15. ROC of EapGAFS

4.2 Diabetes Dataset

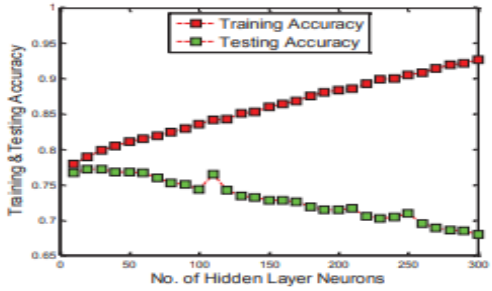


Fig.16.Training and testing accuracy of ELM

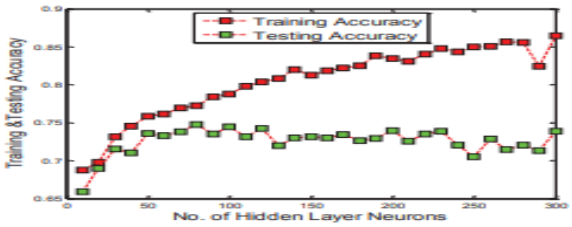


Fig.17.Training and testing accuracy of Adaboost

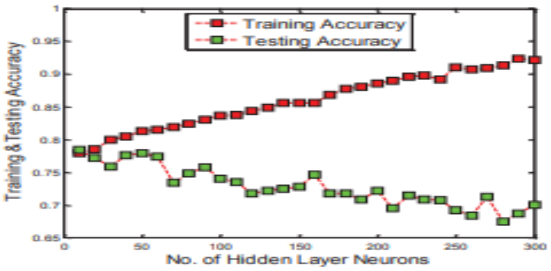


Fig.18.Training and testing with EapGAFS

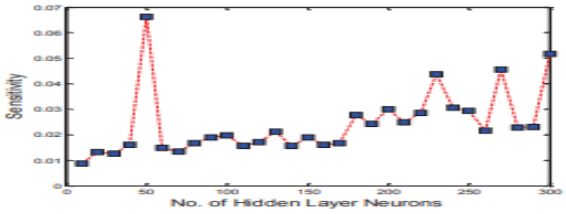


Fig.19. Sensitivity of ELM

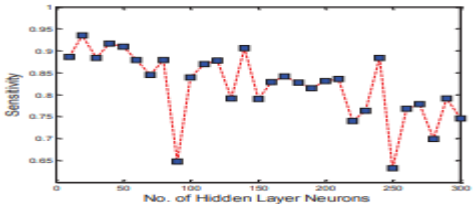


Fig.20: Sensitivity of Adaboost

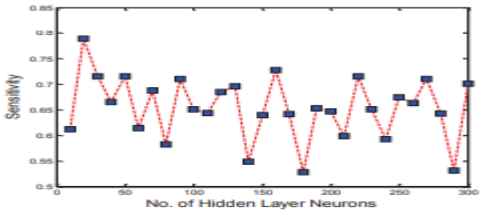


Fig.21.Sensitivity of EapGAFS

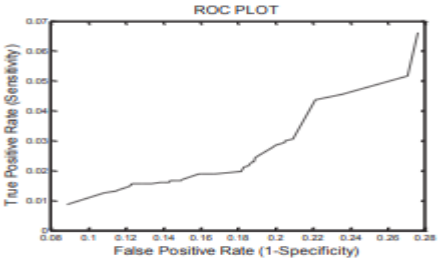


Fig. 22: ROC for ELM

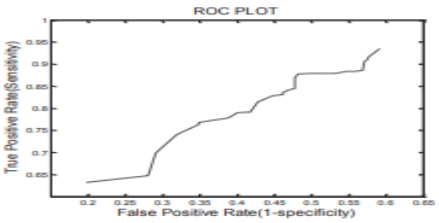


Fig.23: ROC for Adaboost

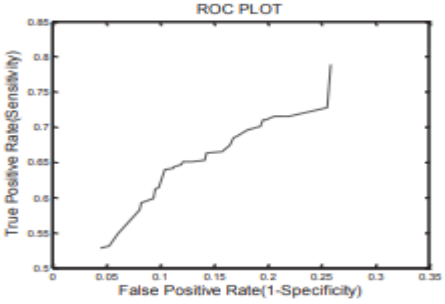
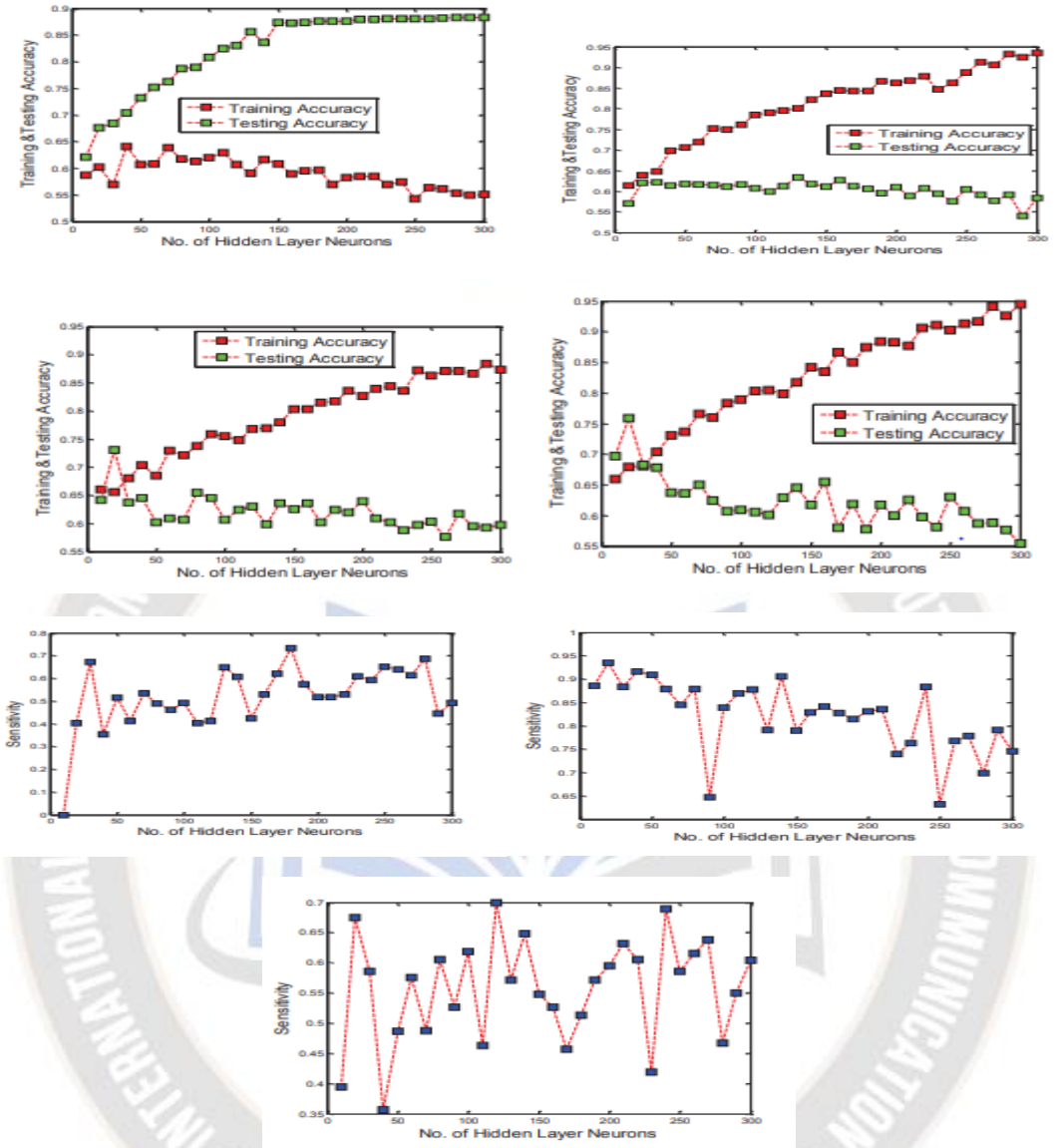


Fig.24: ROC for EapGAFS

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v10i8.5664
***Article Received**: 10 June 2022 **Revised**: 15 July 2022 **Accepted**: 30 July 2022 **Publication**: 31 August 2022*

4.3 Bupa Dataset



The computation is based on consideration of patient data collected from a sample patient count of around 60k. The generated confusion matrix provides the TP, TN, FP, and FN values. The performance of the proposed EapGaFS is comparatively examined with AdaBoost and ensemble classifier. The values obtained with the proposed EapGAFS for the proposed model is presented in table 2 long with the comparison of existing classifiers.

Table 2. Comparison of Classifier

Parameters	AdaBoost	Ensemble	EapGAFS
TN	41439	42098	41702
FN	1109	1596	756
FP	840	181	577
TP	21518	21031	21871

In figure 11 and figure 12 based on epochs, corresponding loss and accuracy values are estimated for training and testing. The proposed EapGAFS incorporates 80% of data for training and 20% for testing.

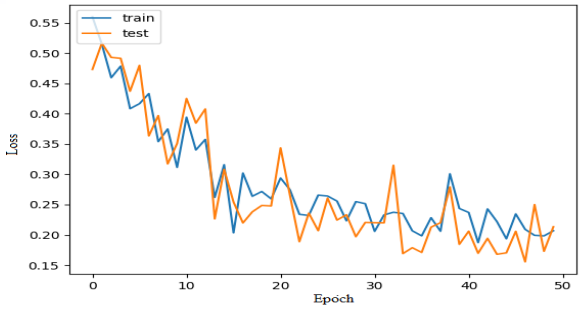


Fig.28. Computation of Accuracy

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v10i8.5664
***Article Received**: 10 June 2022 **Revised**: 15 July 2022 **Accepted**: 30 July 2022 **Publication**: 31 August 2022*

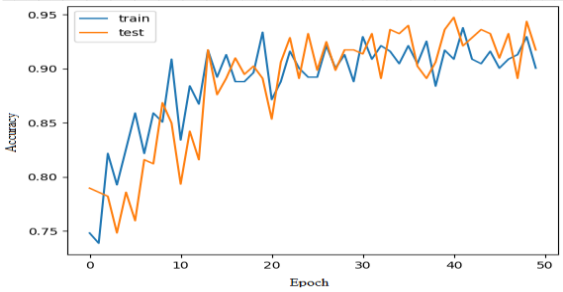


Fig.27. Computation of Loss

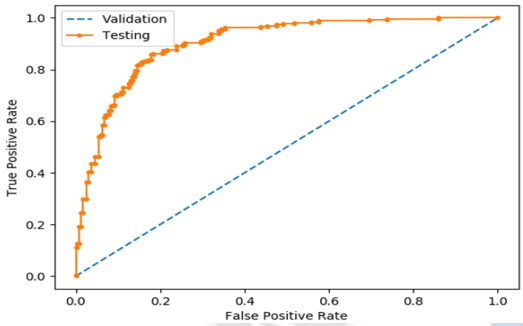


Fig.29. Computation of ROC Curve

Figure 29 provides the ROC curve value based on the testing and validation for e-healthcare applications. The performance of the proposed EapGAFS offers improved loss and accuracy for varying numbers of epochs. Based on generated patient dataset comparative analysis is performed for the proposed EapGAFS with existing conventional classifiers such as ANN, SVM, Decision Tree, AdaBoost.

The simulation analysis expressed that the proposed EapGAFS offers significant performance improvement than the conventional classifiers. Based on the generated confusion matrix and performance metrics the parameters computed are presented in table 3.

Table 3. Comparison of Performance

Parameters	AdaBoost	Ensemble	EapGAFS
Accuracy %	97	97	98
Precision %	96	94	97
Recall %	95	93	97
F1 – Score	0.955	0.96	0.975

The comparative analysis of performance metrics of the classifiers expressed that the proposed EapGAFS exhibits superior performance than the conventional classifiers such as AdaBoost, conventional ANN, and ensemble. The accuracy measurement of the existing classifier AdaBoost, ANN, and ensemble provides the accuracy value of 97, 96, and 97, whereas the proposed EapGAFS achieves the accuracy value of 98%. In the computation of precision proposed EapGAFS achieves the value of 97 which is significantly higher than the conventional classifiers. In the computation of recall, the existing classifier Adaboost, ANN, and ensemble classifier provide the value of 95, 94, and 93, where the proposed EapGAFS offers a recall value of 97%. Similarly, the proposed EapGAFS achieves the F1 – score value of 0.975 which is significantly higher than the existing classifiers. The analysis expressed that the proposed EapGAFS classifier provides an improved performance than the existing classifiers of ~2 – 3%.

5. Conclusion

This paper concentrated on microarray dataset classification using the ensemble classifier integrated with the GA. This paper developed a ensemble model integrated with a GA algorithm for the classification of microarray dataset. The proposed EapGAFS collects the microarray dataset for sources and uses the GA for clustering. The proposed EapGAFS scheme uses the GA algorithm integrated with the Aprior learning for the data clustering in microarray dataset. Through collected data, analysis is performed for the classification of microarray datset for diseases diagnosis. The analysis of results expressed that the proposed EapGAFS achieves improved performance than the conventional classifiers such as AdaBoost, and ensemble classifiers. The analysis of results expressed that the proposed EapGAFS achieves a superior performance of ~ 4 – 6% than the existing AdaBoost and ensemble classifier. In future, the proposed model can be further extended with the multi-objective optimization model for performance improvement.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Alanni, R., Hou, J., Azzawi, H., & Xiang, Y. A, “novel gene selection algorithm for cancer classification using microarray datasets,” BMC medical genomics, vol.12,no.1,pp.1-12.2019.

[2] Ghazaly, N. M. . (2022). Data Catalogue Approaches, Implementation and Adoption: A Study of Purpose of Data Catalogue. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 8(1), 01–04. https://doi.org/10.17762/ijfrcsce.v8i1.2063

[3] Cilia, N. D., De Stefano, C., Fontanella, F., Raimondo, S., & Scotto di Freca, A,“An experimental comparison of feature-selection and classification methods for microarray datasets,”Information, vol.10,no.03, pp.109,2019.

[4] Basavegowda, H. S., & Dagnew, G,“Deep learning approach for microarray cancer data classification,”CAAI Trans. Intell. Technol.,vol.5,no.1,pp.22-33,2020.

[5] Zeebaree, D. Q., Haron, H., & Abdulazeez, A. M,“Gene selection and classification of microarray data using convolutional neural network,”In 2018 International Conference on Advanced Science and Engineering (ICOASE), Duhok, Iraq, 2018,pp. 145-150.

[6] Ahmed Cherif Megri, Sameer Hamoush, Ismail Zayd Megri, Yao Yu. (2021). Advanced Manufacturing Online STEM Education Pipeline for Early-College and High School Students. Journal of Online Engineering Education, 12(2), 01–06. Retrieved from http://onlineengineeringeducation.com/index.php/joee/article/view/47

[7] Aydadenta, H., & Adiwijaya, A, "A clustering approach for feature selection in microarray data classification using random forest," Journal of Information Processing Systems, vol.14,no.5,pp.1167-1175,2018.

[8] Zhang, G., Hou, J., Wang, J., Yan, C., & Luo, J, "Feature selection for microarray data classification using hybrid information gain and a modified binary krill herd algorithm," Interdisciplinary Sciences: Computational Life Sciences, vol.12,no.3,pp.288-301,2020.

[9] Potharaju, S. P., & Sreedevi, M, "Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance,"Clinical Epidemiology and Global Health,vol.7,no.2,pp. 171-176,2019.

[10] Ghosh, M., Adhikary, S., Ghosh, K. K., Sardar, A., Begum, S., & Sarkar, R, "Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods,"Medical & biological engineering & computing, vol.57,no.1, pp.159-176,2019.

[11] Mohammed, M., Mwambi, H., Omolo, B., & Elbashir, M. K, "Using stacking ensemble for microarray-based cancer classification," In 2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE) , Khartoum, Sudan,2018,pp. 1-8.

[12] Bai, A., & Hira, S. "Microarray cancer classification using feature extraction-based ensemble learning method,"International Journal of Data Analysis Techniques and Strategies, vol.13,no.3, pp.244-263,2021.

[13] Manikandan, G., & Abirami, S, "A survey on feature selection and extraction techniques for high-dimensional microarray datasets,"In Knowledge computing and its applications Springer, Singapore,2018, pp. 311-333.

[14] Sun, M., Liu, K., Wu, Q., Hong, Q., Wang, B., & Zhang, H, "A novel ECOC algorithm for multiclass microarray data classification based on data complexity analysis,"Pattern Recognition, vol.90, pp.346-362,2019.

[15] Das, A. K., Pati, S. K., & Ghosh, A, "Relevant feature selection and ensemble classifier design using bi-objective genetic algorithm,"Knowledge and Information Systems, vol.62,no.2,pp.423-455,2020.

[16] Sayed, S., Nassef, M., Badr, A., & Farag, I, "A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets,"Expert Systems with Applications, vol.121, pp.233-243,2019.

[17] Hengpraprohm, S., Hengpraprohm, K., Thammasiri, D., & Mukviboonchai, S, "Co-evolving ensemble of genetic algorithm classifier for cancer microarray data classification.,"Advanced Science Letters,vol.24,no.2, pp.1330-1333,2018.

[18] Shukla, A. K, "Multi-population adaptive genetic algorithm for selection of microarray biomarkers,"Neural Computing and Applications, vol.32,no.15,pp.11897-11918,2020.

[19] Barnali, S., Satchidananda, D., & Kumar, J. A."Usage of ensemble model and genetic algorithm in pipeline for feature selection from cancer microarray data," International Journal of Bioinformatics Research and Applications,vol.16,no.3, pp.217-244,2020.

[20] Talatian Azad, S., Ahmadi, G., & Rezaeipanah, A.,"An intelligent ensemble classification method based on multi-layer perceptron neural network and evolutionary algorithms for breast cancer diagnosis,"Journal of Experimental & Theoretical Artificial Intelligence, pp.1-21,2021.

[21] Gangavarapu, T., & Patil, N, "A novel filter–wrapper hybrid greedy ensemble approach optimized using the genetic algorithm to reduce the dimensionality of high-dimensional biomedical datasets,"Applied Soft Computing, vol.81,pp.105538,2019.

[22] Sayed, S., Nassef, M., Badr, A., & Farag, I,"A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets,"Expert Systems with Applications, vol.121, pp.233-243,2019.

[23] Jena, J. J., & Satapathy, S. C, "Use of Evolutionary Algorithms for Detection of Fatal Diseases via DNA Micro-array Classification: A Review,"Communication Software and Networks, Springer, Singapore ,2021,pp.649-656.

[24] Patra, B., Jena, L., Bhutia, S., & Nayak, S, "Evolutionary Hybrid Feature Selection for Cancer Diagnosis,"In Intelligent and Cloud Computing ,Springer, Singapore,2021. pp. 279-287.

[25] Bagheri Khoulenjani, N., Saniee Abadeh, M., Sarbazi-Azad, S., & Jaddi, N. S, "Cancer miRNA biomarkers classification using a new representation algorithm and evolutionary deep learning,"Soft Computing, vol.25,no.4,pp.3113-3129,2021.

[26] Anwer, K. I., & Servi, S. (2021). Clustering Method Based on Artificial Algae Algorithm. International Journal of Intelligent Systems and Applications in Engineering, 9(4), 136–151. https://doi.org/10.18201/ijisae.2021473632

[27] Abdulaal, A. H., Shah, A. F. M. S., & Pathan, A.-S. K. (2022). NM-LEACH: A Novel Modified LEACH Protocol to Improve Performance in WSN. International Journal of Communication Networks and Information Security (IJCNIS), 14(1).

[28] Debata, P. P., & Mohapatra, P, "Selection of informative genes from high-dimensional cancerous data employing an improvised meta-heuristic algorithm,"Evolutionary Intelligence,pp.1-19,2021.

[29] Talatian Azad, S., Ahmadi, G., & Rezaeipanah, A, "An intelligent ensemble classification method based on multi-layer perceptron neural network and evolutionary algorithms for breast cancer diagnosis," Journal of Experimental & Theoretical Artificial Intelligence, pp.1-21,2021.

[30] Taspinar, Y. S., Koklu, M., & Altin, M. (2021). Fire Detection in Images Using Framework Based on Image Processing, Motion Detection and Convolutional Neural Network. International Journal of Intelligent Systems and Applications in Engineering, 9(4), 171–177. https://doi.org/10.18201/ijisae.2021473636

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v10i8.5664
***Article Received**: 10 June 2022 **Revised**: 15 July 2022 **Accepted**: 30 July 2022 **Publication**: 31 August 2022*

[31] Hamarsheh, Q., Daoud, O. R., Al-Akaidi, M., Damati, ahlam, & Bani Younis, M. (2022). Robust Vehicular Communications Using the Fast-Frequency-Hopping-OFDM Technology and the MIMO Spatial Multiplexing. International Journal of Communication Networks and Information Security (IJCNIS), 14(1).

[32] Gumaei, A., Sammouda, R., Al-Rakhami, M., AlSalman, H., & El-Zaart, A, "Feature selection with ensemble learning for prostate cancer diagnosis from microarray gene expression.,"Health Informatics Journal,vol.27,no.(1),pp.1460458221989402,2021.

[33] Almugren, N., & Alshamlan, H, "A survey on hybrid feature selection methods in microarray gene expression data for cancer classification," IEEE access, vol.7,pp.78533-78548,2019.