Efficient Machine Learning Technique for Tumor Classification Based on Gene Expression Data

C. Venkatesan[1]
Department of ECE,
HKBK College of Engineering,
Bangalore, Karnataka, India
venkatintphd@gmail.com

T.Thamaraimanalan[2]
Department ofECE,
Sri Eshwar College of Engineering,
Coimbatore, Tamilnadu, India

D.Balamurugan[3]
Department of CSE,
Sona College of Technology, Salem,
Tamilnadu, India

M.Ramkumar[4]
Department of CSE,
HKBK College of Engineering,
Bangalore, Karnataka, India

***Abstract*: In bioinformatics research, cancer classification is a crucial domain. The use of microarray technology to identify specific illnesses is common. A small number of genes uncovered in clinical applications can lead to low-cost medicines that can help estimate a patient's survival time or diagnose cancer. Because there are more genes and fewer samples in microarray data, high dimensionality is a serious concern. The genes in the microarray data were evaluated using F-statistics,T-Statistics, and Signal-to-Noise Ratio (SNR) in this study.The top-m rated genes are analyzed using optimization approaches to retrieve useful information. The genetic algorithm (GA), particle swarm optimization (PSO), cuckoo search (CS), and shuffling frog leaping with rapid flying are among the methods employed (SFLLF). Classification is done using the Support vector machine (SVM), the K-Nearest Neighbor classifier (KNN), and the Naive Bayes classifier (NBC). Lung Cancer Michigan, AML-ALL, Colon Tumour, Lung Harvard2, and others are among the datasets utilized for experimental analysis. The classifiers are assessed using a 5-fold cross-validation approach. The findings demonstrate that the suggested two-step feature selection approaches are effective in selecting relevant genes from microarray data for cancer classification.**

***Keywords:Signal to noise ratio (SNR), Genetic algorithm (GA), Shuffled Frog Leaping with Levy Flight (SFLLF), Naive Bayes Classifier (NBC*).**

I.INTRODUCTION

Cells are the elementary structure of all living beings. Humans will have trillions of cells, and each cell contains an exact copy of the genome which is encoded in Deoxyribonucleic acid. The hereditary material in humans and all other organisms is DNA. Almost all of the cells in the human body share the same DNA. The cell nucleus contains the bulk of DNA, while the mitochondria contain just a little amount. Chromosomes are long DNA molecules organized into chromosomes in cells. A gene is a chromosomal component that specifies how to construct a protein[1-2].

The method by which the data contained in a gene is turned into a phenotype (protein)that can be observed is known as gene expression. The presence of Messenger ribonucleic acid (mRNA) in a tissue indicates that a gene is active in that tissue. As a result, changes in gene expression can cause cell death or uncontrolled growth, such as cancer [3].

II.MICROARRAY TECHNOLOGY

Bioinformatics is the application of informatics technologies to biological and biomedical datasets to improve knowledge discovery in both biology and computer science [4]. Microarray technology has been increasingly popular in biomedical research in recent years. A DNA microarray is a solid-surface array of tiny DNA patches.It provides functional relationship information at the genome level between cellular and physiological processes of biological organisms and genes. Variations in gene expression levels were found to be related to the risk of developing a disease. Figure 1 depicts the microarray technology process.

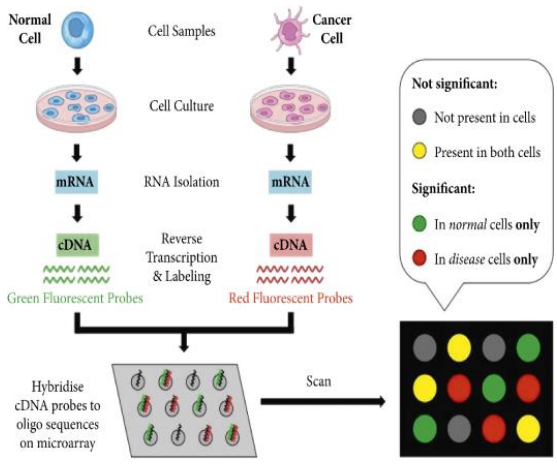

Fig 1: Microarray technology

1. Problem formulation

Finding relevant genes or gene combinations with high predictive efficacy from microarray data for cancer classification is the most difficult problem in bioinformatics. The large dimensionality of gene expression data is a major drawback. There are more genes in it, but there are fewer samples. To choose useful genes for illness prediction and diagnosis, feature selection approaches are required. The challenge of feature selection from a huge dataset is NP-

978-1-6654-0814-1/22/$31.00 ©2022 IEEE

hard(non-deterministic polynomial-time). To increase classification accuracy [5], methods based on feature or gene choice eliminate inappropriatefeatures. It requires determining a subset of the original characteristics so that a classifier based on this subset outperforms a classifier based on the complete set [6,7].

Most existing microarray-based cancer classification approaches, on the other hand, rely on far too many genes to accomplish reliable classification. Because many classification algorithms are not scalable to high dimensions, they are inapplicable to analyzing raw gene microarray data [8]. The purpose of classification is to discover which genes are differentially expressed and may be used to predict class membership in future data.The complications are as follows:

- Dimension reduction to reduce the computational cost
- Reduction of noisy and irrelevant genes to improve classifier performance
- The selection of significant genes that can aid in the identification and monitoring of the target diseases

This study uses statistical measures and heuristic optimization techniques for feature selection to address the aforementioned challenges. Initially, statistical measures such as T-Statistics, SNR, and F-Statistics were applied to microarray data to rank genes based on their discrimination capability, which was then used as a gene marker. To identify relevant genes from the top-m ranked genes, heuristic optimization approaches such as the Genetic algorithm(GA), Particle swarm optimization (PSO), Cuckoo Search (CS), and Shuffled frog leaping with levy flight (SFLLF) are utilized. KNN, SVM, and NBC were used as classifiers. To analyze the effectiveness of the suggested feature selection methods [8], the accuracy of the classifier was assessed.

2. Proposed GA Based Feature Selection for Cancer Classification

Computing approaches based on biological evolution's processes and mechanisms are known as evolutionary algorithms. They're comparable in that they adapt through an iterative process of trial and error that builds up and strengthens beneficial variation. Members of a population struggling to live in a problem-specific objective function-defined environment are candidates for solutions.A sort of algorithm used to solve optimization problems is called an evolutionary algorithm (GA). In the proposed method it provides an overview of GA and how it is utilized in cancer classification feature selection, as well as how to interpret experimental results.

3. Simple Genetic Algorithm

A computer technique for solving optimization issues is based on biological evolution. The Genetic algorithm (GA) is an optimization algorithm based on probability. The population is the basis for this algorithm. In a stochastic approach, natural selection principles are applied to build successive populations of possible solutions. GA, in particular, models the biological mechanisms that allow a population's subsequent generations to adapt to their circumstances. Adaptation is largely accomplished by genetic transmission from parents to children, with the survival of the fittest acting as the driving principle. It also collects data and determines the length of time the evolution should endure.

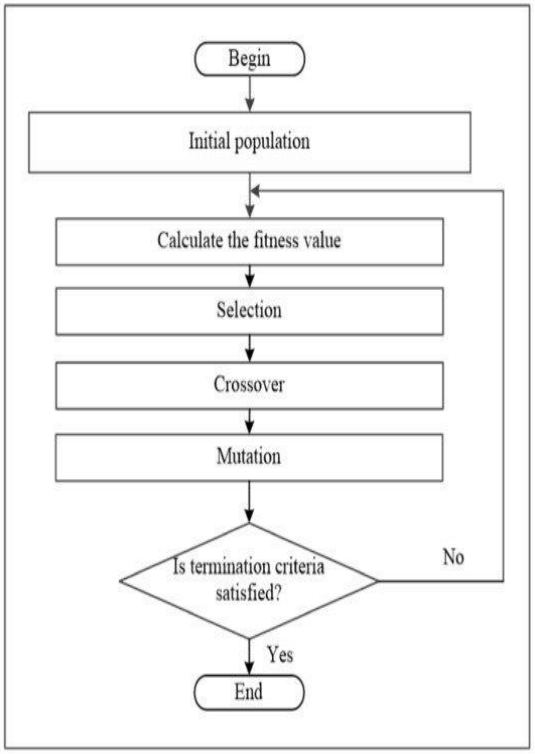

Fig 2: Flowchart of the proposed algorithm

Fewer fit individuals are likely to be chosen for reproduction. To create children, a pair of parents from the current generation's intermediate population was randomly selected to mate. To boost the un-predictable structure, the mutation operator is employed to modify one or more genes on a randomly chosen chromosome. Finally, a different form of the selection process is utilized to duplicate and transfer the individuals who survive from one generation to the next. To use GA to tackle a specific problem, follow these steps.

- Represent chromosomal encoding or alternate solutions.
- Choose the fitness function for the best solution.
- Discover the suitable genetic operators.

III. RESULTS AND DISCUSSION

A good setup can lead the algorithm to quickly converge on the best results, but a bad configuration leads to taking a long time to discover a decent solution. Population size, crossover rate, mutation rate, and generation number are the four primary factors of GA. SGA will be used in this analysis. A roulette-wheel selecting mechanism was employed during the replication process. A crossover is introduced with a probability of Pc. In GA applications, a crossover rate of 0.65 to 1 is advantageous since mutation happens with such a probability for each bit of the string. The amount of the population in Georgia is quite important.

The magnitude of the population in GA is crucial. It influences the size of the memory as well as the pace of

convergence. An increase in population size or generations can be used to expand the search space.

As a result, crossover probability is 0.65, the chances of mutationare 1/L, the population size is 100, and the maximum number of iterations is 400; the chromosome length 'L' might be anywhere between 10 and 100 years old. The GA parameters and values are listed in Table 1 below.

TABLE 1:GA PARAMETERS AND THEIRVALUES

GA parameters	Values
Choice type	Roulette wheel selection
Crossovertype	One-point
Mutationtype	Flip bit
Evaluation type	Elitists
Population size	100
Quantity of generations	400
Selection ratio	0.4
Crossover ratio	0.65
Chances of a mutation	1/L

For the Ovarian Cancer, Lung Harvard2 datasets, DLBCL Harvard, Lung Cancer Michigan,AML-ALL dataset, SVM and, the suggested feature selection strategy achieves 100% classification accuracy with other classifiers. The classification accuracy of all statistical metrics and classifiers for the CNS dataset is 81.25 percent. The classification accuracy of all classifiers for the DLBCL outcome dataset and the Prostate outcome dataset is 77.27 percent and 85.71 percent, respectively. For the Colon Tumor dataset, the SNR feature selection strategy provides 96 percent classification accuracy with other classifiers. For all classifiers with F-statistics values, the greatest classification accuracy attained for the Prostate dataset is 93.86 percent. Table 2 compares the classification accuracies obtained using GA and other classifiers.

Table 3 shows the outcomes of the GA-based feature selection approach for the considered dataset, as well as comparisons to other published methods.

TABLE 2:CLASSIFICATION ACCURACY COMPARISON ACHIEVED WITH VARIOUS CLASSIFIERS USING GA

Dataset	KNN Classifier	SVM Classifier	NBC Classifier
CNS	84.50+~#	84.25+~#	84.25+~#
DLBCL Harvard	98.25~	98.25+~	98.25~
DLBCL outcome	87.42+~	84.27+~#	88.36+
Lung Cancer	100.00~#	100.00+~#	100.00~#
Ovarian Cancer	98.75#	98.75~#	98.75#
Prostate outcome	88.51+~	88.51~#	88.51+~#
AML-ALL	96.54~#	96.25~#	96.54~#
Colon Tumor	96.00~	96.00~	96.00~
Lung Harvard2	91.45~#	91.45#	91.75~#
Prostate	93.86#	93.86#	93.86#

+ T-Statistics; ~ SNR; # F- Statistics

TABLE 3:GA CLASSIFICATION ACCURACY COMPARED TO OTHER APPROACHES

Reference / Dataset	CNS	DLBCL Harvard	DLBCL outcome	Lung Cancer	Ovarian Cancer	Prostate Outcome	AML-ALL	Colon Tumor	Lung Harvard2	Prostate
[1]	-	-	-	-	-	-	-	90.68	-	-
[3]	82.54	96.00	77.45	96.00	97.00	-	97.5	89.41	99.63	-
[11]	100	-	-	-	100	-	100	100	-	-
Proposed work	83.25	96.00	77.45	96.00	98.00	85.71	100	92.00	99.63	92.68

TABLE 4:GENESWERE CHOSEN FOR THE AML-ALL DATASET USING A GA-BASED FEATURE SELECTION APPROACH

Gene Name	Analysis id	Gene Report
TRA@	M21624	T cell receptor alpha locus
CNN1	D17408	Calponin 1, smooth muscle
CCR5	U83326	Chemokine (C-C motif) receptor 5
ELA2	M27783	Neutrophil, Elastase 2,
IFNA6	X02958	Alpha 6, Interferon,
MAP3K3	U78876	Mitogen-activated protein kinase 3
CLDN10	U89916	Claudin 10
TNNI1	J04760	Troponin I type 1
DAZL	U66726	Azoospermia
PTGDR	U31099	Prostaglandin D2 receptor (DP)
NCBP1	D32002	Nuclear cap binding protein 1
PIK3R2	J03909	Phosphoinositide-3-kinase, regulatory 2

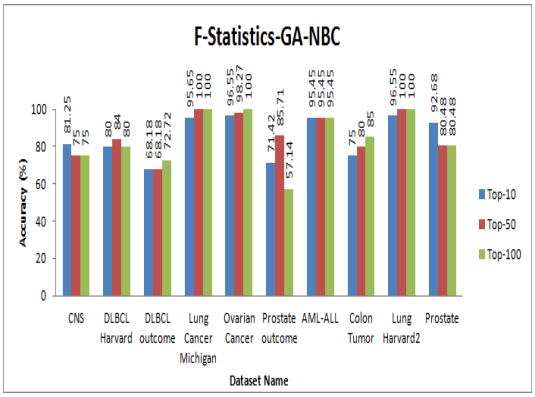

Fig3:F-statistics and GA with NBC improved classification accuracy

The list of genes picked for the AML-ALL dataset using a GA-based feature selection approach is shown in Table 4. Figure 2 displays F-statistics and GA with NBC improved classification accuracy. Figures 3 and 4 show the accuracy of classification using F statistics and GA with SVM and KNN classifiers.

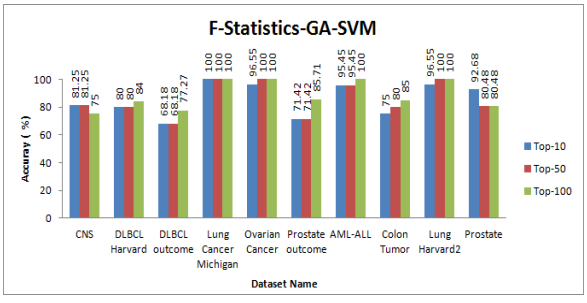

Fig4:F-statistics and GA with SVM both improve classification accuracy

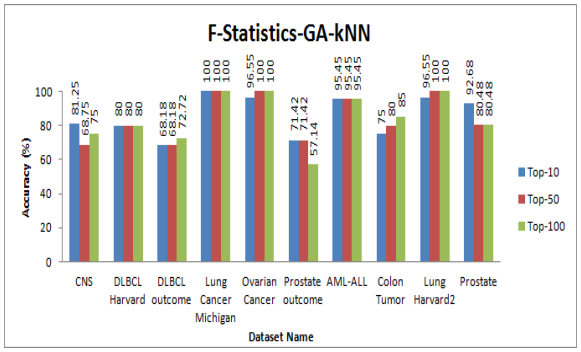

Fig 5:F-statistics and GA with KNN improve classification accuracy

CNN1, PIK3R2, TRA@, LCK, and MAP3K3 were selected as important genes using GA-based feature selection. The PIK3R2 gene codes for a human enzyme called phosphatidylinositol 3-kinase regulatory subunit beta. The TRA@ gene codes for the protein T-cell receptor alpha locus (TRA@), which is discovered in humans. LCK is a 56-kilodalton protein found inside lymphocytes, which are immune system cells. A protein is coded for by the gene MAP3K3. The stress-activated protein kinase is directly controlled by this protein, which activates SEK and MEK1/2. AML-ALL cancer can be caused by these genes, according to research.

IV. CONCLUSION

GA, like biological evolution, generates a solution through the notion. Different parameter values must be set to improve GA outcomes. For 5 of 10 cancer datasets, including Lung Harvard2, AML-ALL, Ovarian Cancer, Lung Cancer Michigan, DLBCL Harvard, the GA-based feature selection approach achieves 100 percent classification accuracy. When it comesto categorizing data, GA believes that SNR and F-statistics feature selection techniques outperform F-Statistics. SVM performs better than both KNN and NBC. The GA-assisted feature selection technique provides better results than the statistical measure-based strategy when it comes to classification accuracy. The features picked by GA for various cancer datasets have been discovered to be biologically significant. The results show that increasing the number of genes involved in the classifier does not improve its performance.

REFERENCES

[1] Arevalillo, J. M., & Navarro, H. (2013). Exploring correlations in gene expression microarray data for maximum predictive–minimum redundancy biomarker selection and classification. Computers in Biology and Medicine, 43(10), 1437–1443. https://doi.org/10.1016/j.compbiomed.2013.07.005

[2] Bennet, M., Akiva, A., Faivre, D., Malkinson, G., Yaniv, K., Abdelilah-Seyfried, S., Fratzl, P., &Masic, A. (2014). Simultaneous Raman microspectroscopy and fluorescence imaging of bone mineralization in living zebrafish larvae. Biophysical Journal, 106(4). https://doi.org/10.1016/j.bpj.2014.01.002

[3] Chen, K.-H., Wang, K.-J., Tsai, M.-L., Wang, K.-M., Adrian, A. M., Cheng, W.-C., Yang, T.-S., Teng, N.-C., Tan, K.-P., & Chang, K.-S. (2014). Gene selection for cancer identification: A decision tree model empowered by particle swarm optimization algorithm. *BMC Bioinformatics*, *15*(1). https://doi.org/10.1186/1471-2105-15-49

[4] Chuang, L.-Y., Yang, C.-S., Wu, K.-C., & Yang, C.-H. (2011). Gene selection and classification using Taguchi chaotic binary particle swarm optimization. Expert Systems with Applications, 38(10),13367–13377. https://doi.org/10.1016/j.eswa.2011.04.165

[5] Thamaraimanalan, T., RA, L., & RM, K. (2021). Multi biometric authentication using SVM and ANN classifiers. Irish Interdisciplinary Journal of Science & Research (IIJSR).

[6] LI, J.-T., & JIA, Y.-M. (2010). An improved elastic net for cancer classification and gene selection. Acta AutomaticaSinica, 36(7), 976–981. https://doi.org/10.3724/sp.j.1004.2010.00976

[7] Nabeel, M., Rehman, A., & Shoaib, U. (2017). Accuracy based feature ranking metric for multi-label text classification. International Journal of Advanced Computer Science and Applications, 8(10). https://doi.org/10.14569/ijacsa.2017.081048

[8] Roth, G. A., Forouzanfar, M. H., Moran, A. E., Barber, R., Nguyen, G., Feigin, V. L., Naghavi, M., Mensah, G. A., & Murray, C. J. L. (2015). Demographic and epidemiologic drivers of Global Cardiovascular Mortality. New England Journal of Medicine, 372(14), 1333–1341. https://doi.org/10.1056/nejmoa1406656