Genetic Algorithms and Feature Selection for Improving the Classification Performance in Healthcare

Alaa Alassaf[1]*, Eman Alarbeed[2], Ghady Alrasheed[3], Abdulsalam Almirdasie[4],
Shahd Almutairi[5], Mohammed Abullah Al-Hagery[6], Faisal Saeed[7]
Department of Computer Science-College of Computer, Qassim University, Qassim, Kingdom of Saudi Arabia[1, 2, 3, 4, 5, 6]
College of Computing and Digital Technology, Birmingham City University, Birmingham, United Kingdom[7]

***Abstract*—Microarray technology appeared recently and is used in genetic research to study gene expressions. Microarray has been widely applied to many fields, especially the health sector, such as diagnosing and predicting diseases, specifically cancer diseases. These experiments usually generate a huge amount of gene expression data with analytical and computational complexities. Therefore, feature selection techniques and different classifications help solve these problems by eliminating irrelevant and redundant features. This paper presents a proposed method for classifying the data using eight classifications machine learning algorithms. Then, the Genetic Algorithm (GA) is applied to improve the selection of the best features and parameters for the model. We use the higher accuracy of the model among the different classifications as a measure of fit in the genetic algorithm; this means that the model's accuracy can be used to select the best solutions than others in the community. The proposed method was applied to the colon, breast, prostate, and Central Nervous System (CNS) diseases and experimental outcomes demonstrated an accuracy rate of 93.75, 96.15, 82.76, and 93.33 respectively. Based on these findings, the proposed method works well and effectively.**

***Keywords*—Cancer classification; gene expression; feature selection; microarray data; algorithm; machine learning; genetic algorithm**

I. Introduction

In particular, the study of gene expression data has significant implications for diagnosing and treating different diseases, including cancer. This is because an organism's traits and characteristics are defined by its genes, which are the basic building blocks of heredity. About 20,000–25,000 genes in humans are responsible for different aspects of growth and development. The instructions to create a specific protein are encoded in a Deoxyribonucleic acid (DNA) sequence known as a gene. Mutations in the gene sequence can cause protein structure or function changes, leading to genetic diseases and disorders.

Recent developments in gene expression analysis have made it possible for researchers to study the levels of gene activity in specific cells or tissues, shedding light on the diseases' underlying causes. The classification of gene expression data is essential in bioinformatics research since it may be used in several applications. Some of these applications are to find possible biomarkers for disease diagnosis and treatment. Several techniques, like Chi Square and Support Vector Machine (SVM) with Recursive Feature Elimination (RFE), have been put forth to classify gene expression data and show promise to perform so accurately. Both methods have been previously used for gene expression data classification, with varying degrees of success. For instance, [1] used the ChiSquare method and SVM for gene expression classification and reported an accuracy of 89.57%. Similarly, [2], used SVM-RFE and other machine learning algorithms for gene expression data classification and reported an accuracy of 92.5%.

A recent study [3], proposed a feature selection method called (ChiSVMRFE). That combines the Chi-squared test and SVMs to identify a subset of features most informative for classification.

Several studies have previously compared different feature selection and classification methods for gene expression data, including [4] – [7].

Motivated to advance biological knowledge discovery from gene expression profiles, we aim to comprehensively evaluate feature selection and classification combinations applied to multiple cancer datasets. Specifically, we seek to:

- Evaluate technique performance using microarray data on prostate, colon, CNS, and breast cancer in a systematic way. Precision in disease diagnosis and treatment may be greatly impacted by this finding.
- Find the best performing integrated feature selection-classification techniques. addressing problems such as small sample sizes and class imbalance, addressing problems such as small sample sizes and class imbalance that genetic algorithms can help address.
- Gain new knowledge to direct the search for biomarkers by conducting comprehensive analyses.

Random Forest, Logistic Regression, KNN, SVM and Decision Tree are applied as classifiers. A GA conducts feature selection to optimize informative genes. Integrating selection with diverse classifiers addresses gene expression challenges while capturing different patterns.

A GA conducts feature selection to optimize informative genes while addressing challenges in microarray data analysis.

*Corresponding Author.
The Deanship of Scientific Research, Qassim University

Integrating selection with diverse classifiers aims to comprehensively analyze datasets through leveraging their individual strengths.

Substantial effort compiled the breast, colon, CNS and prostate datasets from thousands of genes, warranting comprehensive evaluation. Therefore, the contribution of this paper includes the following:

- Developing a framework to a hybrid genetic algorithm-classifier.
- Enhancing the results by extensive dataset analysis.
- Achieving greater accuracy compared to earlier efforts.

This is how the rest of the paper is structured. Section II briefly presents the literature review related to Gene Expression, High-dimensional Problems, Feature Selection Methods, and Classification. Section III presents the methodology, the datasets used in this paper, and the preprocessing techniques employed on the data. In contrast, Section IV explains the experiments, whereas Section V highlights the results and discussion. Finally, Section VI includes the conclusions and future works.

II. Literature Review

A. Gene Expression

Cancer research is one of the major areas of research in the medical field. Cancer is a group of related diseases with a high mortality rate characterized by abnormal cell growth, which attacks the body tissues [3], [8] – [10]. Microarray cancer data is a prominent research topic across many disciplines focused on addressing problems related to the higher curse of dimensionality, a small number of samples, noisy data, and imbalance class [11]– [17]. The Microarray technology allowed the researchers to analyze thousands of gene expression profiles relevant to different fields, including medicine, especially cancer [13] – [15], [18]. With the rapid improvement of DNA microarray tools and technology, researchers can simultaneously measure hundreds of genes expression levels [3].

Gene expression profiling uses microarray techniques to discover gene patterns when expressed. However, because microarrays produce a large volume of data, the analysis procedure requires a lot of computation power and time [14].

B. High-Dimensionality Problem

Gene expression datasets with high dimensionality consist of a large number of genes and a small number of samples. In classification problems based on the microarray, the data usually contains many irrelevant and redundant features [19]. Various approaches have been used to solve high-dimensional problems and predict the most required features within limited datasets. Usually, the used technique of high-dimensionality is called least absolute shrinkage and selection operator (LASSO), which is one of the main concepts in dealing with high-dimensional cancer classification [11], to choose the best subset of features for microarray data. A gene selection approach [19], eliminates duplicated and unnecessary characteristics to pick the optimal subset of features for microarray data. In study [20], a novel hybrid instance learning-based filter wrapper approach addresses a high-dimensionality issue in which a small sample size is transformed into a tool that enables selecting a small number of feature subsets which has proven effective. A proposed model [11], called the Adapted Penalized Logistic Regression (CBPLR) model, uses the total number of selected genes, the Area Under the Curve (AUC), and the misclassification rate (i.e., error rate). It is evaluated on three popular high-dimensional cancer classification datasets, which shows how effective the model is for classifying cancer.

In addition, [16], an approach called Shapely Value Embedded (SVEGA) is proposed, which increases the accuracy of breast cancer detection by selecting the gene subset from the high-dimensional gene data. Four classifiers distinguish between normal and abnormal tissues to identify benign and malignant tumors. As a result, classification accuracy shows that the proposed approach leads to a better breast cancer diagnosis and greater performance.

C. Feature Selection Methods

To diagnose cancer in human bodies, the feature selection method is a search problem among various genes for an optimal solution that detects the most gene expressed [8]. The gene selection strategy eliminates duplicated and unnecessary characteristics to pick the optimal subset of features for microarray data [19], [15]. In most cases, there are a lot of genes but not many samples in gene expression data. for this reason, traditional gene selection based on mutual information using machine learning models has data sparseness problems [10].

Machine learning algorithms have called the attention of researchers due to their ability for pattern recognition in data [8] [10]. Likewise, [22], provides two selection approaches for SVMRFE-based discriminative feature subsets for measuring the feature subset. Techniques such as SVM-RFE address this by combining classification accuracy and sample overlapping measures to accurately assess feature subsets. Furthermore, an experimental study has employed the Markov Blanket-Embedded Genetic Algorithm (MBEGA) [21]. It is successful and it provides the best balance among all four assessment criteria: accuracy, number of genes, computational cost, and robustness.

Additionally, surveys like the one mentioned [18] provide valuable insights into the taxonomy of feature selection methods, highlighting challenges such as high dimensionality and unbalanced classes.

As a result, new techniques keep emerging yearly, not limited to improving previous approaches' classification accuracy results.

D. Classification

In cancer classification, various approaches for measuring gene expression, such as Fisher's linear discriminant analysis, nearest neighbor analysis, and max-margin classifiers. Despite advancements, challenges like computation time, classification accuracy, and biological relevance persist [23]. In current microarray technology, feature reduction is critical and sensitive in the classification task to achieve satisfactory classification accuracy [18], [24]. One of the main barriers to technology adoption is the analysis and management of such data [13]– [15], [18], [25].

One of the solutions is the SVM a popular and efficient classification technique widely applied in many fields, especially biological [3], [9], [10], [21], [25]. This can be combined with RFE to be SVM-RFE, which is used for an efficient feature selection technique that is based on SVM and increases Classification effectiveness [3], [17], [22]. In [12], the SVM approach has been used with the Leave-One Out Cross-Validation (LOOCV) approach (i.e., reserving the trained data point while it trains the rest of the dataset) to classify genes effectively. In study [8], classification strategy makes use of memetic algorithms, which speed up the entire evolutionary searching process by introducing a Local Search (LS) operation. Specifically, that is algorithm starts from a candidate solution. Next, makes minor perturbations while moving to a neighboring solution then, the process is repeated until a solution deemed optimal is found.

A hybrid cancer classification approach involving several machines learning tools, including Pearson’s correlation coefficient, decision tree classifier, and cross-validation (CV) to optimize the maximum depth hyperparameter. The result shows that the model improves classification accuracy [11]. In [26], a three-phase hybrid approach has been used to select and classify high dimensional microarray data. It combines several classifiers via Pearson Correlation Coefficient (PCC), Binary Particle Swarm Optimization (BPSO), or GA which shows improved classification accuracy.

In research [13], a combination of supervised and unsupervised data analysis including the categorization of cancer and the prediction of gene function classes. It goes through how potential regulatory signals in the genomic sequences may be predicted using the gene expression matrix, and then it explores several potential directions for the future. As a result, it shows that analysis methods of gene expression data will advance and become more organized.

III. METHODOLOGY

A. Datasets

In this study, we leverage four high-dimensional microarray datasets, each corresponding to a distinct type of cancer: Breast cancer, Colon cancer, Central Nervous System (CNS) cancer, and Prostate Cancer. These datasets are pivotal for understanding the complex gene expression profiles associated with each cancer type and for identifying potential biomarkers for diagnosis and treatment strategies [3].

As shown in Table I, the key characteristics of each microarray dataset are:

1) Breast cancer dataset: Comprises a comprehensive set of 16,382 features covering 36,626 genes, classified into two categories. This dataset is instrumental in studying the genetic variations specific to breast cancer, aiding in the identification of unique gene expression patterns.

2) Colon cancer dataset: Contains 2,000 features for 2,000 genes, all categorized into two groups. This dataset allows for the exploration of genetic factors that contribute to colon cancer, facilitating the development of targeted therapies.

3) Prostate cancer dataset: Includes 12,646 features for 12,646 genes, with the data divided into two classes. This dataset provides insights into the genetic underpinnings of prostate cancer, offering opportunities for discovering novel genetic markers.

4) CNS cancer dataset: Features 7,129 features for 7,129 genes, organized into two classes. This dataset is crucial for unraveling the genetic complexity of CNS cancers, potentially leading to breakthroughs in understanding the disease's molecular basis.

Preprocessing steps are customized for each dataset to accommodate varying data types, including normalization for continuous features, and encoding for categorical variables, ensuring data uniformity and integrity for the feature selection process.

TABLE I. MICROARRAY DATASETS DESCRIPTION

Datasets	Feature	Genes	Classes
Breast	16382	36626	2
Colon	2000	2000	2
Prostate	12646	12646	2
CNS	7129	7129	2

B. Feature Selection Methods

Feature selection is a helpful preprocessing method to decrease the data dimensions and enhance classification accuracy [23]. Selecting groups of useful genes with high prediction potential from current samples is one of the numerous issues in bioinformatics. The abnormally high dimension of the search space presents the biggest challenge in analyzing gene expression data. The most common way to display gene expression data is in a matrix with many genes and a few samples. Its objective is to eliminate properties that don’t help with the classification issue or are redundant because they offer the same data. Finding pertinent genes for subpopulation samples is the first step in the feature selection process for cancer data in microarray data. In a binary category data collection, the sample is typically categorized as having either cancer or not having cancer [8].

C. Genetic Algorithm

A GA is an evolutionary algorithm that is a metaheuristic-inspired natural selection process. It starts by producing a random beginning population. In this technique, GA operators, which include selection, interception, and mutation, are used to search for the best solutions by individuals [26]. The survival of the fittest member of a population that changes over time is the central tenet of the genetic algorithm. The population is first evaluated and initialized. A fitness function that assesses the effectiveness of the problem-solving solution evaluates each individual. Through generations, the GA iterates to create changes in the population. Three evolutionary operators are applied to the population once every generation.

The first operator is the selection operator, which picks a group of people to keep in the following generation or, more appropriately, to be merged with again by the other operators. Natural selection directly influences the operator of this fitter, for people are more likely to be chosen. Crossover is the second operator used on the population, which involves taking advantage of the shared space between two people the selection operator has chosen. It combines the two people, the parents, to create the two new people, the children. The final operator, mutation, randomly alters a person’s genes to broaden the population’s genetic diversity. The mutation rate is typically chosen at a modest value since many mutations could cause the GA to devolve into a simple random search. The population evolves until the stop condition is reached. At this point, the best estimate for a particular problem is returned [8].

D. *Classifiers*

In this study, we use several classifiers to classify the data, including Random Forest, Logistic, KNeighbors, Gradient Boosting, LinearSVM, RadialSVM, AdaBoost, and DecisionTree.

E. *Optimization of Classifier Parameters*

In parallel to feature selection, our GA approach extends to optimizing classifier hyper parameters. By encoding hyper parameters as part of the chromosomes, we ensure that each feature subset is evaluated using the best possible classifier configuration, thereby maximizing classification performance.

IV. Experiments

The method used in this study is based on a GA to identify features in a cancer dataset. This method improves the statistical performance of machine learning models by removing unnecessary features from the dataset. It includes feature encoding, population generation, intersection, and final feature selection. A GA provides an efficient way to select the best features in a data set, as it is used to generate a set of potential solutions and then select the solutions that perform best based on the given performance metric.

A. *Implementation of GA for Feature Selection*

As shown in Fig. 1, the GA starts with a random set of chromosomes. Each chromosome represents a potential solution to the feature selection problem. The fitness function is then applied to the chromosomes. The chromosomes with the highest fitness scores are then selected for reproduction. The chromosomes of the offspring are then created using crossover and mutation factors. The offspring’s chromosomes are then evaluated and repeated until the stopping criterion is met. The discontinuation criterion can be based on the number of generations, chromosomal fitness, or a combination. The GA is a powerful tool for feature selection. It can be used to find optimal solutions to the feature selection problem in various fields.

B. *Function Description for Genetic Algorithm*

1) Functions for splitting the dataset:

a) Split (): This function splits the dataset into training and testing sets. It takes the dataset and the ratio of the training set as inputs and returns the training set and the testing set.

2) Functions for evaluating the classifiers:

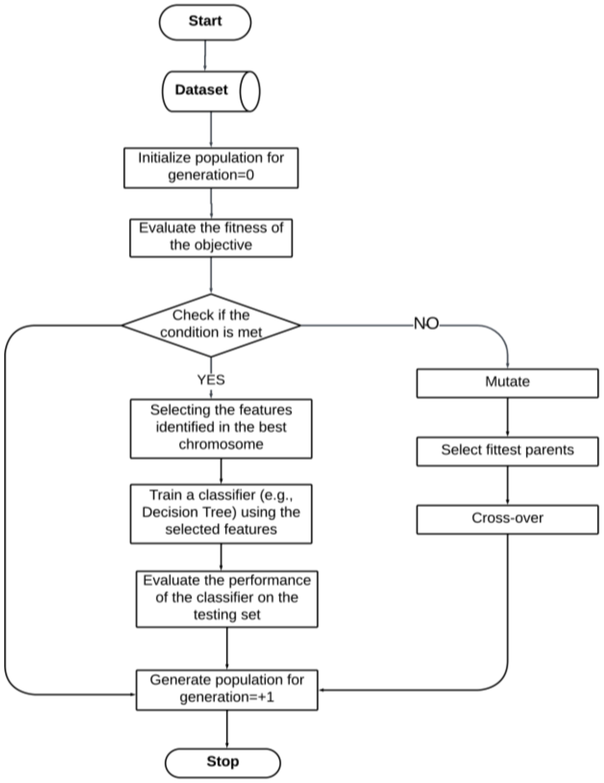
Start
Dataset
Initialize population for generation=0
Evaluate the fitness of the objective
Check if the condition is met
NO
YES
Mutate
Select fittest parents
Cross-over
Selecting the features identified in the best chromosome
Train a classifier (e.g., Decision Tree) using the selected features
Evaluate the performance of the classifier on the testing set
Generate population for generation=+1
Stop

Fig. 1. Flowchart of GA application steps to choose the best features.

a) Acc-score (): This function is used to evaluate the performance of multiple classifiers on the dataset. It takes the training set, testing set, and a list of classifiers as input and returns the accuracy score of each classifier on the testing set.

3) Functions for plotting:

a) Plot (): This function is used to plot the results of the genetic algorithm. It takes the number of generations and the best fitness score of each generation as input and plots a line graph.

4) Functions for the genetic algorithm:

a) Initialization-of-population (): This function generates an initial population of chromosomes for the genetic algorithm. It takes the number of features and the population size as input and returns a randomly generated population of chromosomes.

b) Fitness-score (): This function calculates the fitness scores of the chromosomes. It takes the population, training, and testing set as input and returns the best parents and their fitness scores.

c) Selection (): This function selects the best parents for the next generation. It inputs the training and res and returns the best parents.

d) Crossover (): This function performs crossover between the best parents to generate offspring. It takes the best parents as input and returns offspring chromosomes.

e) Mutation (): This function introduces new genetic variation in the offspring chromosomes. It takes the offspring chromosomes, and the mutation rate as input and returns mutated offspring chromosomes.

5) Generations (): This function executes all the above functions for the specified number of generations. It takes the population, the training set, the testing set, the number of generations, the number of features, and the mutation rate as input and returns the best chromosome (the set of selected features) and its fitness score.

C. Implementation Steps

1) Reading dataset from a CSV file.

2) Splitting the data into sets for testing and training: The dataset is split into training and testing sets. The training set is used to train the classification model, while the testing set is used to evaluate its performance.

3) Encoding the classes in the training set into numerical values: The target class of each sample in the training set is encoded into a numerical value.

4) Creating the fitness function: The fitness function represents the performance of a selected feature set (chromosome) in a classification task using various classifiers, such as decision trees, K-nearest Neighbors (KNN), and other techniques, using the selected features. For example, the fitness function can be defined as the accuracy of a decision tree classifier using the selected features.

5) Creating the objective function: The objective function determines the direction of the search for the optimal solution. The objective function is to maximize fitness function.

6) Specifying the initial size of the chromosome population: The initial population size of the chromosomes (sets of selected features) is specified.

7) Generating a random set of chromosomes: A random set of chromosomes is generated to start the genetic algorithm.

8) Computing fitness scores for each chromosome: The fitness function is applied to each chromosome in the population, and the fitness score is computed.

9) Selecting the best chromosomes and using them to produce new generations: The best chromosomes in the population are selected to produce new generations of chromosomes. In this example, the selection process is based on the fitness scores of the chromosomes.

10) Creating new generations using crossover and mutation factors: The new generations of chromosomes are created by applying crossover and mutation operators. Crossover involves exchanging the selected features between two chromosomes, while mutation involves randomly changing a selected feature in a chromosome.

11) Computing fitness scores for the new chromosomes: The fitness function is applied to the new chromosomes, and the fitness score is computed.

12) Selecting the best chromosomes in the new generations: The best chromosomes in the new generations are selected. In this example, the best chromosomes have the highest fitness scores.

13) Until the stopping requirement is satisfied, repeat steps 10–12: Steps 10-12 are repeated until a stopping criterion is met. In this example, the stopping criterion is a fixed number of iterations.

14) Selecting the features in the best chromosome: The features selected in the best chromosome are identified as the optimal set of features for the classification task.

15) Training a classification model using the selected features: A classification model (e.g., KNN classifier) is trained using the optimal set of features.

16) Evaluating the model's performance using the testing set and computing the accuracy and confusion matrix: The performance of the trained classification model is evaluated using the testing set, and the accuracy and confusion matrix is computed.

V. Experimental Results and Discussion

This section presents and discusses the results of the conducted experiments. The Comparisons between the obtained results and other studies are also presented in this section. The proposed method for selecting features with both discrete and continuous values based on the application of the GA was tested to optimize the accuracy after being classified using several classifications on four databases.

A. Experimental Results

1) Classifiers with original datasets: After applying the classifications to the data for the four diseases of the colon, Breast, Prostate, and CNS. We find that the accuracy is variable and different for diseases and the type of classification. The accuracy of classification models depends on a combination of factors, including the nature of the disease, data complexity, dataset size and quality, choice of features, and classification algorithm type are presented in Table II. The experimental results are summarized as follows:

a) The performance of several different classification algorithms was compared on a set of original data for four different datasets: Colon, Breast, Prostate, and CNS.

b) The Random Forest algorithm showed strong performance, recording the highest accuracy rate in the Colon dataset at 81.25% and in the breast dataset at 76.92%.

c) The K-Neighbors algorithm performed exceptionally well in the Colon dataset with an accuracy rate of 87.50%, indicating its effectiveness in handling this dataset.

d) Linear SVM demonstrated good performance in the CNS dataset with an accuracy rate of 80%, while achieving acceptable results in the Breast and Prostate datasets with accuracy rates of 76.92% and 68.97%, respectively.

e) AdaBoost showed good accuracy in the Breast dataset at 80.77%, but its performance was lower in the Prostate dataset, where it reached 62.07% accuracy.

f) The Decision Tree and Gradient Boosting algorithms performed well in the Colon dataset with an accuracy rate of

84.62% but did not achieve the same level of success in the other datasets.

g) Radial SVM had the lowest performance in most datasets, with low accuracy rates, especially in the breast dataset, where accuracy was 46.15%.

TABLE II. THE ACCURACY OF APPLYING THE CLASSIFIERS ON THE ORIGINAL

Classification	Dataset			
	Colon	Breast	Prostate	CNS
Random Forest	81.25	76.92	**75.86**	73.33
Logistic	75	76.92	68.97	73.33
K-Neighbors	**87.50**	73.08	72.41	46.67
Linear SVM	62.50	76.92	68.97	**80**
Radial SVM	81.25	46.15	51.72	66.67
AdaBoost	81.25	80.77	62.07	53.33
Decision Tree	62.50	**84.62**	51.72	73.33
Gradient Boosting	68.75	**84.62**	68.97	73.33

2) Genetic algorithm: In this way, after applying the different classifications to the four diseases of the colon, breast, prostate, and CNS. We used the higher accuracy of the model among the different classifications as a measure of fit in the genetic algorithm. This means that the model's accuracy can be used to determine which solutions are better than others in the population. Then, the GA is applied to improve the selection of the best features and parameters of the model from the data set, and its steps are as follows:

a) Split data: The split () function used in training and test data in the GA is used to evaluate the model's performance.

b) Initialization of the population: This is done by creating a Boolean array of size n-feat, where n-feat is the number of features in the dataset. The array's first int(size*n-feat) elements are set to False, and the remaining elements are set to True. The population is then shuffled randomly.

c) Fitness Evaluation: This is performed by fitting the model with the data using the Boolean array and then calculating the accuracy score using the model. Predict () method.

d) Selection: This is performed by selecting the n-parents' best chromosomes from the population used for that population-NextGen. append(pop-after-fit(i)).

e) Crossover: This is performed by randomly selecting two parent chromosomes and then performing crossover to generate two new offspring chromosomes.

f) Mutation: This is performed by randomly selecting a chromosome from the population and then performing a mutation to generate a new chromosome at a rate of 0.20.

g) Repeat steps 3-6 until the desired number of generations is reached (The algorithm stops when the best score in the last generation is the same as the previous generations).

h) Return the best score and the corresponding chromosome.

The accuracy score within a generation is determined by the accuracy of the predictions made by the population within that generation. This is calculated using data from the population and labels. Five generations have been produced, and the highest model accuracy is typically equivalent or slightly superior to the accuracy of the preceding generation. Therefore, we use the model's accuracy to compare which values are more valid across generations. Table III and Fig. 2 display the optimal score for four diseases across the first to fifth generations. It is important to note that the numbers 1-5 denote the generation number, with 1 representing the first generation, 2 representing the second generation, and so on. This allows for a clear and concise comparison of model accuracy across multiple generations.

TABLE III. BEST SCORE IN GENERATIONS FOR FOUR DISEASES AFTER APPLYING

Dataset	Generation				
NO. of generation	1	2	3	4	5
Colon	87.5	**93.75**	**93.75**	**93.75**	**93.75**
Breast	92.31	**96.15**	**96.15**	92.31	92.31
Prostate	79.31	79.31	**82**.76	79.31	79.31
CNS	80	80	80	86.67	**93.33**

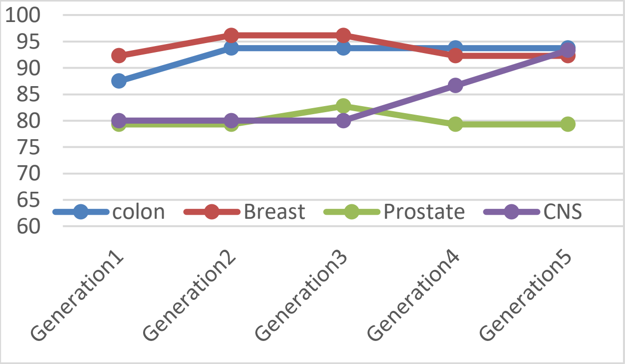

Fig. 2. The best score in the generations for the four diseases.

- The GA showed an improvement in feature selection. It works by simulating the process of natural selection, where solutions with better fitness scores are selected and used to create the next generation of solutions. High accuracy rates were obtained, particularly in the colon, breast, and CNS.
- Decreased accuracy in prostate disease because of Imbalanced data is a common problem in many medical datasets, including prostate cancer data. In addition, insufficient data on the size of the dataset can affect the accuracy.
- The accuracy of the GA is variable because of the randomness of the mutation and selection processes. The mutation process randomly changes the features of the chromosome, and the selection process randomly selects the chromosomes for the next generation. This

means that the model’s accuracy can vary from generation to generation.

- The result of a GA can change each time the code is run because the algorithm uses randomness to generate new solutions and evaluate them. Therefore, the criterion for measuring the accuracy of the GA was the higher accuracy of the classifications.

B. Analysis and Discussion

1) Comparison of Data Accuracy Before and After GA Utilization: Table IV illustrates a comparison of data accuracy before and after using the GA for performance enhancement. A noticeable increase in accuracy was observed after employing GA in all datasets. In the Colon dataset, accuracy improved from 87.50% before using GA to 93.75% after its use. For the Breast dataset, accuracy increased from 84.62% to 96.15% because of GA. In the Prostate dataset, performance was enhanced from 75.86% to 82.76% with GA. As for the CNS dataset, a significant accuracy boost was observed, increasing from 80% before GA to 93.33% after its application. Table IV signifies the effectiveness of GA in enhancing the performance of statistical models for classifying specific diseases.

TABLE IV. COMPARISON OF ACCURACY: ORIGINAL METHOD WITH PROPOSED METHOD

Dataset	Accuracy Using the Original Method	Accuracy Using the Proposed Approach
Colon	87.50	**93.75**
Breast	84.62	**96.15**
Prostate	75.86	**82.76**
CNS	80	**93.33**

2) Compared with the Chi-SVM-RFE method: The proposed method showed improved accuracy for the colon, Breast, and CNS, which achieved high accuracy achieved high accuracy results after applying eight classifications except prostate, the accuracy rate has gone down. Decreased classification accuracy for prostate cancer is due to several factors that can affect the accuracy of classification algorithms on medical data, such as the quality and quantity of the data and the pre-processing and normalization techniques used. Moreover, we assume in this study that the choice of hardware and software used for implementation can also affect the accuracy of classification methods. For example, different devices may have different processing speeds, memory capacities, and computational architectures that can affect the performance of machine learning algorithms (e.g., Random Forest). In addition, the underlying operating systems and software libraries may have different versions and configurations that can affect the run time behavior and accuracy of the models. In Table V, the proposed study was compared with previous studies using the Chi-SVM-RFE method [3] for colon, breast, prostate, and CNS diseases.

TABLE V. COMPARISON OF THE PROPOSED METHOD WITH PREVIOUS STUDIES

Dataset	Proposed	Chi-SVM-RFE
Colon	93.75	**95.24**
Breast	**96.15**	94
Prostate	82.76	**96.09**
CNS	**93.33**	88.33

In Table V, we note that the proposed method achieves high accuracy in the results compared with the previous method, except for the colon and prostate, where there is a clear difference. The reason may be attributed to the fact that the Chi-SVMRFE, or RFE with Support Vector Machines and a Chi-squared criterion, is a statistical approach that aims to eliminate the least iteratively informative features from the dataset. Genetic algorithms use a randomized process of mutation and selection to optimize solutions to problems. Both approaches have their pros and cons, and which one to choose depends on various factors, including the specific type of cancer, the size and quality of the dataset, and the specific research question being addressed.

VI. CONCLUSION

Feature selection has a vital role in preprocessing, especially regarding large data volumes such as cancer microarray data, helps reduce the dimensions of microarrays and improves classification accuracy. The study's contribution was to improve the performance of cancer disease classifications based on the data set collected for cancer diseases. Where a method was applied using eight classifiers, which are Random Forest, Logistic, K-Neighbors, Linear SVM, Radial SVM, AdaBoost, Decision Tree, and gradient boosting based on Datasets of four diseases, it includes including colon, breast, prostate, and CNS. The GA was applied to five generations. The best accuracy for each generation was by measuring its suitability with the highest accuracy of the model among the different classifications. It achieved excellent and high results with breast cancer, reaching an accuracy of 96.15.

On the other hand, the GA showed the lowest accuracy results with the prostate dataset due to insufficient population size. The reason is that the GA is based on the diversity of the population to explore the search space and find the optimal solution. The GA was compared with previous methods ChiSVM-RFE, which showed improvements in breast and CNS datasets. Feature selection is an exciting area of research in multiple fields, such as data mining, pattern recognition, machine learning, statistics, bioinformatics, and genomics.

Therefore, this research contributes to helping to identify the genome to diagnose and understand diseases such as cancer. Early detection may also help predict them but extends to finding the appropriate treatment.

On the other hand, the deficiency in the performance of the proposed method appears only with some types of cancer, such as prostate cancer, because cancer classification is inherently complex due to the heterogeneous nature of the cancer itself. While genetic algorithms may be powerful improvement tools, their effectiveness in classifying cancer depends on various aspects, including problem complexity, quality and characteristics of the data set, and appropriate tuning of algorithm parameters.

Regarding future research, the scope of work can be expanded to apply the GA by improving the quality of the fitness function, the selection criteria, and the population size by generating more generations and a higher mutation rate. In addition, this work can be extended by using the same methodology and mixed feature selection methods.

ACKNOWLEDGMENT

The authors would like to thank the Deanship of Scientific Research, Qassim University, for funding the publication of this project.

REFERENCES

[1] Y. Li and Y. Li, “A novel gene expression data classification method with improved chi-square feature selection and SVM,” *Journal of Biomedical Informatics*, vol. 87, pp. 28–36, 2018.

[2] R. Tabares-Soto, S. Orozco-Arias, V. Romero-Cano, V. S. Bucheli, J. L. Rodr´ıguez-Sotelo, and C. F. Jimenez-Var on, “A comparative study of´ machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data,” *PeerJ Computer Science*, vol. 6, APR 13, 2020, [Online; accessed 2023-05-24].

[3] T. Almutiri and F. Saeed, “Chi-square and support vector machine with recursive feature elimination for gene expression data classification,” in *2019 First International Conference of Intelligent Computing and Engineering (ICOICE)*. IEEE, 2019, pp. 1–6.

[4] X. Chen, Y. Li, and Y. Chen, “A comparative study of feature selection methods for gene expression data classification,” *Computational and Mathematical Methods in Medicine*, vol. 2019, pp. 1–10, 2019.

[5] A. J. Ruano-Sanchez and I. Rodr´ ´ıguez-Fdez, “A comparison of feature selection methods for gene expression data classification,” *Journal of Biomedical Informatics*, vol. 71, pp. 145–156, 2017.

[6] L. Wang, G. Zhu, and Y. Guo, “A comparative study of feature selection and classification methods for gene expression data,” *BioMed Research International*, vol. 2016, pp. 1–11, 2016.

[7] F. M. Khan, Y. Liu, and M. F. Ijaz, “A comparative study of machine learning algorithms for gene expression data classification,” *PloS one*, vol. 15, no. 3, p. e0229858, 2020.

[8] M. G. Rojas, A. C. Olivera, J. A. Carballido, and P. J. Vidal, “Memetic micro-genetic algorithms for cancer data classification,” *Intelligent Systems with Applications*, vol. 17, 2 2023.

[9] Z. Zhu, Y. S. Ong, and M. Dash, “Markov blanket-embedded genetic algorithm for gene selection,” *Pattern Recognition*, vol. 40, pp. 3236–3248, 11 2007.

[10] G. Dagnew and B. H. Shekar, “Ensemble learning-based classification of microarray cancer data on tree-based features,” *Cognitive Computation and Systems*, vol. 3, pp. 48–60, 3 2021.

[11] Z. Y. Algamal and M. H. Lee, “Penalized logistic regression with the adaptive lasso for gene selection in high-dimensional cancer classification,” *Expert Systems with Applications*, vol. 42, pp. 9326–9332, 12 2015.

[12] C. D. A. Vanitha, D. Devaraj, and i. Venkatesulu, “Gene expression data classification using support vector machine and mutual information-based gene selection,” *Procedia computer science*, vol. 47, pp. 13–21, 2015.

[13] A. Brazma and J. Vilo, “Gene expression data analysis,” *FEBS letters*, vol. 480, no. 1, pp. 17–24, 2000.

[14] T. Almutiri, F. Saeed, M. Alassaf, and E. A. Hezzam, “A fusion-based feature selection framework for microarray data classification,” pp. 565–576, 2021.

[15] R. K. Singh and M. Sivabalakrishnan, “Feature selection of gene expression data for cancer classification: A review,” vol. 50. Elsevier B.V., 2015, pp. 52–57.

[16] S. Sasikala, S. A. Balamurugan, and S. Geetha, “A novel feature selection technique for improved survivability diagnosis of breast cancer,” vol. 50. Elsevier B.V., 2015, pp. 16–23.

[17] L. Sun, X. Zhang, Y. Qian, J. Xu, and S. Zhang, “Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification,” *Information Sciences*, vol. 502, pp. 18–41, 10 2019.

[18] M. A. Hambali, T. O. Oladele, and K. S. Adewole, “Microarray cancer feature selection: Review, challenges and research directions,” pp. 78–97, 6 2020.

[19] S. Guo, D. Guo, L. Chen, and Q. Jiang, “A centroid-based gene selection method for microarray data classification,” *Journal of Theoretical Biology*, vol. 400, pp. 32–41, 7 2016.

[20] A. B. Brahim and M. Limam, “A hybrid feature selection method based on instance learning and cooperative subset search,” *Pattern Recognition Letters*, vol. 69, pp. 28–34, 2016.

[21] S. Zhu, D. Wang, K. Yu, T. Li, and Y. Gong, “Feature selection for gene expression using model-based entropy,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, pp. 25–36, 1 2010.

[22] X. Lin, C. Li, Y. Zhang, B. Su, M. Fan, and H. Wei, “Selecting feature subsets based on svm-rfe and the overlapping ratio with applications in bioinformatics,” *Molecules*, vol. 23, 2018.

[23] Y. Lu and J. Han, “Cancer classification using gene expression data,” pp. 243–268, 2003.

[24] H. Fathi, H. Alsalman, A. Gumaei, I. I. Manhrawy, A. G. Hussien, and P. El-Kafrawy, “An efficient cancer classification model using microarray and high-dimensional data,” *Computational Intelligence and Neuroscience*, vol. 2021, 2021.

[25] Y. Chen, L. Wang, L. Li, H. Zhang, and Z. Yuan, “Informative gene selection and the direct classification of tumors based on relative simplicity,” *BMC Bioinformatics*, vol. 17, 1 2016.

[26] S. S. Hameed, F. F. Muhammad, R. Hassan, and F. Saeed, “Gene selection and classification in microarray datasets using a hybrid approach of PCC-bpso/ga with multi classifiers,” *Journal of Computer Science*, vol. 14, pp. 868–880, 2018.