Transactions on Machine Intelligence 7(3) (2024) 223-237

ISSN Online: 2821-1693

Transactions on Machine Intelligence

Journal Homepage: https://tmachineintelligence.ir/

An Outlier Detection Approach To Highlight Effective Genes By A Deep Learning Model and An Adjusted Genetic Algorithm (DLAGA)

Y. Aliakbarpoor[1] , E. Parvinnia[2,*], S. Setayesh[3]

[1] Department of Computer Engineering, Shiraz Branch, Islamic Azad University, Shiraz, Iran.
[2] Department of Computer Engineering, Shiraz Branch, Islamic Azad University, Shiraz, Iran.
[3] Department of Computer Engineering, Shiraz Branch, Islamic Azad University, Shiraz, Iran.

ARTICLE INFO

Article History:
Received 1 July 2024
Received in revised form 28 August 2024
Accepted 23 September 2024
Available online 29 September 2024

Keywords:
Anomaly Detection, Gene Analysis, Deep Learning, Microarray Dataset, Gene Selection

ABSTRACT

Identifying abnormally expressed genes is a critical step in cancer diagnosis and has attracted significant attention within the biomedical research community. Gene expression datasets typically involve high-dimensional data, which poses major challenges during the pre-processing stage, particularly in maintaining the biological relevance and interpretability of selected genes. Traditional gene selection techniques often struggle with high computational demands and fail to preserve the intrinsic biological meaning of genes. In this study, we present an effective two-phase framework for gene selection and classification tailored for cancer diagnosis. The first phase employs a Variational Autoencoder (VAE), a deep learning-based technique, to reduce data dimensionality while capturing essential gene expression patterns. In the second phase, we utilize an Adjusted Genetic Algorithm (AGA) to search for a subset of informative genes. To further enhance classification performance, we integrate a wrapper-based approach within the AGA to individually classify genes relevant to different cancer types. Our method was evaluated on two publicly available microarray datasets. The experimental results reveal that the proposed framework outperforms several existing approaches in terms of classification accuracy, while maintaining reasonable computational efficiency. The integration of VAE and AGA offers a robust and biologically interpretable approach to gene selection, making it a promising tool for advancing precision oncology. These findings underscore the potential of combining deep learning and evolutionary algorithms for effective biomarker discovery in high-dimensional genomic data.

1. INTRODUCTION

Being the second-leading cause of global mortality, researchers pay great attention to cancer-related topics, particularly from the genetic point of view [1]. In recent years, various studies have been conducted to introduce AI as an excellent tool to facilitate and improve the process of timely diagnosis of diseases. Authors in [2] proved that AI is a useful tool in diagnosing temporomandibular disorders. According to contemporary biological advances, it

* Corresponding author: Elham.parvinnia@iau.ac.ir
Department of Computer Engineering, Shiraz Branch, Islamic Azad University, Shiraz, Iran.

http://dx.doi.org/10.47176/TMI.2024.223

is reported that mutations in genes of human tissues caused abnormality in the gene expression levels, which leads to specific cancer [3]. Targeting the genes mutations in cancer treatment has been discussed recently. Although targeted therapy has dramatically changed treatment outcomes and disease prognosis, in other oncological contexts, targeted approaches should be investigated more carefully [4]. Despite the development of gene expression technology and availability of new datasets at GEO (Gene Expression Omnibus) database, many researchers still tend to use the same old datasets [5].

High-dimensional data refer to those datasets having too many features and few numbers of observations. Having a high dimensional dataset results in notable decline in classification accuracy while enlarging the learning time [6]. In such cases, feature selection as a pivotal player in machine learning and data mining, exerting an intense influence on the performance and interpretability of predictive models [7]. Although in [8] it has been reported that the feature selection techniques include filter, wrapper, and embedded, our extensive research in scientific databases of articles and journals showed us that three other techniques consists of deep learning (DL) models and genetic algorithm as optimization based techniques and in relation with wrapper methods, and hybrid models can be included in this category.

The filter feature selection methods are computationally faster but their performances are not sufficiently accurate and different since the features are evaluated independently of the learning model. In contrast, wrapper methods interact with the learning and as a result it can achieve better results compared to the previous method. However, the wrapper methods are time-consuming in case of applying on high-dimensional gene expression data [9]. In contrast to wrapper methods that train machine learning models and select features sequentially, embedded methods incorporate the feature selection process within the learning or model-building phase. This integration allows for the simultaneous development of machine learning models and the selection of features, resulting in a reduction in computation time [10]. One of the drawbacks of this method is the risk of over-fitting as the embedded methods select the most important features during the training process, which leads to not obtaining desired generalization score [11].

In this article, we have designed and developed a novel approach in two consecutive phases. The first phase's tasks are: to transpose the selected features, and to find the normal distribution of each gene using variational autoencoder (VAE), based on a distribution function. The second phase's duty is to find features even more precisely, which convey further information. Applying an adjusted genetic algorithm (AGA) produces an output consisting of features, or genes that have the largest impact on getting infected by cancer and classifies the effective genes according to cancer type. The second phase is fed by the output of the first phase.

Although the Genetic algorithm suffers from inherent limitation in dealing with high dimensional tasks [12], our mixed method compensates this limitation smartly. In fact before entering to the genetic algorithm, the search space is minimized by usage of a DL purifier, which in our case is VAE.

Deep learning (DL) algorithms are specifically crafted to autonomously acquire knowledge and identify significant patterns and representations from large datasets [13]. VAEs have become increasingly popular in recent years because of their exceptional ability to remove noise, making them valuable for detecting anomalies [14].

It is crucial to acknowledge that neural networks are widely recognized as a groundbreaking and optimal solution, provided that hyper-parameters are appropriately configured rather than being set in a rudimentary manner with random parameters. Grounding on the fact that DL needs reasonable amount of data to be trained accurately, we have imported the features to VAE in a different way. Indeed, to fulfill the DL requirement on having enough number of data and also to calculate the standard deviation and mean of each feature not each sample, we have transposed the selected features matrix.

The rest of this work is organized as follows: Section 2 offers an overview of methods for selecting genes. Section 3 outlines essential concepts used throughout this study for foundational understanding. Section 4 delves into the intricacies of the method we propose. Experimental outcomes are examined in section 5, leading to the final observations and conclusions drawn in section 6.

2. RELATED WORKS

Regarding the issue of finding genes with unusual behavior or generally the issue of gene selection, there are many researches in Google Scholar. We believe researches in this field can be divided into simple and complex categories. Researches that have used pure versions of filter, or wrapper, or embedded methods belong to the first category, and the rest, including deep learning, genetics, and hybrid, belong to the second category. The review of many articles in the field of gene selection, we infer that most of the older researches are associated with the pure filter method and combinations of them. In this section, we review the articles in both categories.

In [15], the authors developed a filter-based feature selection method for temporal gene expression data based on maximum relevance and minimum redundancy criteria. However their approach is greedy and it evaluates each gene in isolation, without considering the interaction between genes. This can lead to suboptimal selection of genes. A comparative study on filter methods such as ReliefF, and Pearson Correlation has been done in [16]. The benefit of these pure methods is simplicity and low cost, however the high accuracy of them is not that much valuable because of not evaluating the selected genes by a classifier.

In wrapper methods, the selection process is wrapped around the model, using its accuracy as the best subset of genes. Although it has better performance for specific tasks compared to filter method, it suffers from the higher computational complexity. A two-step wrapper approach for gene selection is conducted in [17]. This method combines genetic algorithm to limit the number of genes, followed by World Competitive Contests (WCC) to select an optimal subset.

An ensemble filter method followed by genetic optimizer is used in [18]. Although the authors used the union of the results of three pure filters consisting of ReliefF, Chi-Square, and SU, their method not consider the learning model.

In [19], the authors developed an improved version of the Memetic algorithm named RMA to look up the most relevant genes in the microarray. A new local search is employed based on Relief measure and also two new operators added to the basic algorithm. It is reported that it could find the lowest possible number of effective genes with a very high accuracy. Nevertheless, they did not discuss about the execution time, which should be very time consuming according to the high dimensionality of microarray datasets and also the defined recursiveness setting of their method.

In [20], the authors designed a wrapper method called bSCSO, based on the Sand cat swarm optimization. Although they stated the high accuracy and small features size, the datasets they used are not benchmarked.

Authors in [21] introduce an embedded approach that integrates the construction of Knockoff feature genes within a neural network. The method demonstrated improved gene selection effectiveness over other algorithms across various datasets. However, it notes potential overfitting due to high-dimensional gene expression data.

A wrapper-based hybrid model integrating information gain (IG) and Jaya algorithm (JA) for determining the optimum featured genes from high-dimensional microarray datasets is proposed in [22]. Their proposed method has not performed well in terms of accuracy.

In [23], they present an approach to gene selection called the Sine Cosine and Cuckoo Search Algorithm (SCACSA). They checked the accuracy of their proposed method on only one dataset. As they have reported, utilizing advanced neural networks could reveal intricate patterns and enhancing precision in cancer classifications.

Authors in [24] proposed a hybrid quality criterion for gene expression classification by the aim of optimize the hyper parameters of the DL models. In this study no preprocessing step has been seen.

One of the methods included in the wrapper category is the use of clustering techniques and algorithms. In [25], a two-step gene selection method was implemented. Initially, three indices were utilized to filter genes. Subsequently, a genetic algorithm, employing K-means clustering as its fitness function, was developed. A notable drawback of this method is its computational intensity, particularly in calculating the fitness of chromosomes. The inherent complexity of gene interactions within specific cellular pathways challenges the validity of utilizing clustering for gene selection. This complexity is due to the varied expression levels of genes involved in the same pathway, where some may exhibit low expression and others high. Such differential expression, crucial for specific cellular changes, underscores the interaction's significance rather than individual gene expression levels.

Consequently, clustering based on gene expression relative to specific tissues may fail to elucidate the intricate relationships between genes, rendering it an ineffective strategy for discerning gene interactions [26].

Creating a gene network or graph based on expression levels in microarray data and analyzing it using graph theory is a crucial method. Graph-based methodologies have gained prominence, often constructing heterogeneous graphs with diseases and genes represented as nodes and their associations depicted as edges, enhancing the comprehension of complex biological interactions [27]. Despite their effectiveness, graph-based methods face challenges, notably the insufficient interaction among nodes. The efficacy of Graph Convolutional Networks (GCN), as recent research [28, 29] suggest, relies heavily on adequate message passing between nodes to achieve optimal performance. Self-supervised graph structure learning has recently demonstrated effectiveness in retaining diverse knowledge while filtering out irrelevant data. Research in this area primarily explores mutual information estimation techniques to autonomously develop optimal representations of nodes, facilitating a deeper understanding of the underlying graph structure without requiring external supervision [30]. But in general, these techniques have not paid attention to the heterogeneity of external biological networks, and they also require more extensive knowledge in Biology reach reliable outputs.

By considering all these efforts have been done in the gene selection process, we have made the following contributions in this paper:

1. To the best of our knowledge, a hybrid method consisting of VAE deep network and genetic algorithm with our proposed settings on both phases has not been used to select features.
2. Implementation of pre-processing steps appropriate to the problem of anomalous genes detection in microarray datasets.
3. Applying the perspective of outlier detection in analyzing effective genes in causing cancer with the help of a VAE.
4. Benefiting from the appropriate distribution function for better generalization of the problem.
5. Improving the convergence speed of the genetic algorithm by changing the initial population generation policy and using a powerful and fast classifier as the fitness function.

3. BACKGROUND

This section mainly focuses on the preliminaries and requirements that have been used in our proposed method.

3.1. Microarray Data

Microarray technology (MT) is a powerful tool that can simultaneously measure the expression levels of thousands of genes in diverse samples. As a result, it has a major impact on disease management in a variety of applications, from drug discovery to advanced development such as new diagnostic and prognostic tools, to personalized treatment [31]. Microarrays are small chips containing a grid of microscopic spots, each spot with a specific DNA sequence or oligonucleotide. The spots can represent individual genes, allowing analysis of gene expression across different samples and conditions. The process involves: 1) Isolating mRNA from experimental and control samples, 2) Converting the mRNA to complementary DNA (cDNA), 3) Labeling the cDNA with fluorescent dyes, 4) Hybridizing the labeled samples to the microarray chip

After hybridization, the fluorescence level on each spot is measured, reflecting the amount of mRNA bound and the expression level of the corresponding gene. This allows for parallel analysis of gene expression across many genes and samples simultaneously. Due to the simplicity, low cost and high sensitivity of microarrays, this technology is broadly used. [32].

3.2. Variational Autoencoder (VAE)

Variational Autoencoders are built on the principles of autoencoding, where an encoder network learns to compress data into a low-dimensional representation, and a decoder network learns to reconstruct the data from this representation. Unlike standard autoencoders, VAEs introduce a probabilistic twist: they model the latent representation as a distribution, typically Gaussian, from which they sample to generate outputs. This process

introduces variability that can capture the underlying structure of the data more effectively. The VAE consists of two main components:

- **Encoder**: Maps input data x to a latent distribution parameterized by means (μ) and variances (σ^2)
- **Decoder**: Samples from the latent space (z) to reconstruct the input data ($\hat{x}$)

VAEs can be useful for outlier detection because they learn to reconstruct normal data efficiently while struggling to reconstruct outliers. By training a VAE on normal data, the model learns a representation of this data in its latent space. When an outlier is input into the VAE, the reconstruction error (the difference between the input and its reconstruction) tends to be higher than normal data. Thus, by setting a threshold for reconstruction error, outliers can be identified based on their significantly higher reconstruction errors compared to normal data points. VAE not only provides extensive and complex data, but has also been shown to be successful in finding outliers.

3.3. Logistic Regression (LR)

Logistic regression (LR) classifier is categorized as a supervised machine learning algorithm. LR is a robust and straightforward method that accomplishes classification tasks by predicting the probability of an outcome, event, or observation. LR can be used for both binary and multi classes problem.

In binary classification, the outcome is modeled as a function of predictors using the logistic function to ensure the output lies between 0 and 1, interpreted as the probability of the dependent variable belonging to a particular class. This model uses the logit link function,

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n, \quad (1)$$

where p is the probability of the dependent variable being in the positive class. A threshold (commonly 0.5) determines the class assignment. For multi-class problem, a common approach called softmax regression is used.

3.4. Genetic Algorithm(GA)

Genetic Algorithms (GAs) are a class of optimization algorithm inspired by the principles of natural selection and genetics. They simulate the process of natural evolution to solve complex problems by repeatedly selecting, combining, and muting candidate solutions. GAs operates on a population of potential solutions, applying operators such as selection, crossover, and mutation. Through iterative cycles called generations, the algorithm evolves the population toward better solutions based on a fitness function that evaluate their quality. This approach is particularly effective for problems where the search space is large and complex, offering a powerful tool for optimization and problem solving across various scientific and engineering domains.

There are two cornerstones of evolutionary algorithms, which are exploitation and exploration. These two bases need to have an appropriate trade-off in a way that the search process is driven toward global optima with a fast convergence rate [33]. In fact the genetic operators play a pivotal role in a GA success and having a reasonable amount of balance between exploitation and exploration.

Selection operator, which is inspired by natural selection in nature, is used to select a population of chromosomes in each generation. Rank-based, Roulette Wheel and Tournament selection are the most popular selection operator. Another operator of GA called Crossover, which is specific types of recombination, is the main operator of standard GA. As a matter of fact, it is applied to pick a pair of parents to generate new offspring. Simple, Single, and whole are different arithmetic methods to recombine parents when dealing with float numbers. The last operator of the GA is mutation and the mutation techniques are different based on the chromosome representation.

The chromosome representation can be float, binary or permutation. In the problem of anomaly detection in microarray data, both float and binary can be used. In our case we use the float representation to adjust the importance of each gene.

4. PROPOSED METHOD

This section mainly focused on the details of our proposed method. The workflow of our proposed method is shown in figure 2. Considering pivotal role of pre-processing in data any analysis task, we employ a new vision and look at the data from a different perspective. In fact we transpose the dataset as if the features are samples, and vice versa.

4.1. Pre-processing Steps

As it is clear, examining a tissue with microarray technologies can result in tens of thousands of genes for just one sample. Therefore, a small number of samples, along with a large number of features, is one of the characteristics of microarrays [34]. Standardization of microarray datasets guarantess that all variables are on the same scale and have comparable ranges, making them suitable for later analyses [35]. The following steps are executed in sequence:

Robust Scaler: This scaler is specifically designed to handle outliers. Unlike other scaling techniques that are affected by the presence of outliers, RobustScaler uses the median and the interquantile range for scaling, thereby reducing the influence of outliers. This makes it a good choice for anomaly detection problems where outliers are significant concern. Having the hypothesis that we have some genes as outlier in our data, the skewness of the distribution , is inevitable. Despite the use of logarithm two in [26] in the pre-processing process, we do not use this algorithm on the data because the presence of skewness helps more in finding outliers in VAE.

Data Transposition: The final step in the pre-processing pipeline involved the transpostition of the dataset. Initially, the dataset is with samples as rows and genes as columns. For the purposes of our analysis, with focuses on gene-centric examinations, we transposed the dataset such that genes are represented as rows and samples as columns. The reorientation facilitated easeier implementation of subsequent analytical techniques focused on gene expression patterns.

4.2. First Phase of Gene Selection According to Data Distribution (Unsupervised Approach)

We have used a deep learning method with the aim of identifying anomalous features, which actually refer to genes with unusual behavior in the occurrence of diseases. These genes are called outliers. The VAE network has been very successful in identifying this type of data, specially in high-stakes decision making such as medical disgnosis [36]. VAEs operate by benefitting from a distribution function, and is not only rely on the data cumulative frequency. Therefore, with the existence of the Z-Score function in VAE, it can be said that the problem can be generalized. VAE actually identifies genes that do not fit a normal distribution and reports them as outliers. After that, we transpose the dataset again to return to the original form.

4.2.1. Anomaly Detection Using VAE

Following the preprocessing steps, the preprocessed dataset was inputted into a VAE. The VAE was trained to minimize the reconstruction loss and the Kullback-Leibler (KL) divergence, which regularizes the encoder by penalizing deviations of the latent variable distribution from the prior distribution.

Outliers in the microarray dataset were identified based on the reconstruction error and the latent space distribution. Samples with a high reconstruction error were considered outliers, as the VAE struggled to accurately reconstruct these data points, indicating they deviate significantly from the learned data distribution. Additionally, samples that mapped to sparse regions of the latent space were flagged as outliers, under the assumption that normal data points cluster together in the latent representation.

4.2.2. Roles of μ and σ in VAE to Detect Outlier

In VAE the roles of the mean (μ) and standard deviation (σ) are central to its ability to model and generate data, as well as its application in outlier detection. Outliers can often be identified by their high reconstruction error. If the μ and σ for a data point result in a latent representation that is difficult for the decoder to accurately reconstruct, it suggests the data point is not well represented by the model's learned distribution. This is indicative of an outlier.

Analyzing the properties of the latent space can also reveal outliers. Data points whose μ values place them in sparse regions of the latent space, or whose σ values are significantly higher or lower than those of most data points, can be considered outliers. High σ values indicate high uncertainty in how the model represents these data points, possibly because they do not fit well with the majority of the data the model has learned.

4.2.3. VAE Settings in Our Work

We consider 64 neurons for the encoder part. The number of neurons in latent layer of this part is considered to be 2. Then, in the decoder part, 2 neurons for the first layer and then 64 neurons for the next layer are included. Since we do not have negative values in the expression of genes, we use ReLU activation function. We apply Adam

optimizer with the step size (α) equals to 0.001. In addition, we employ the batch normalization and dropout with the value of 0.3 to avoid overfitting.

4.3. Second Phase of Gene Selection Using Genetic Algorithm (Supervised Approach)

After genes with unusual behavior or in other words outliers are identified, we re-transpose the data to return to the original form, so the genes are again in the role of features and not samples. Then we give the genes as input to the second phase consisting of an adjusted genetic algorithm. As a result, in the outlier data search space, we are looking for genes that show the highest amount of abnormal expression and finally we identify them according to the type of cancer.

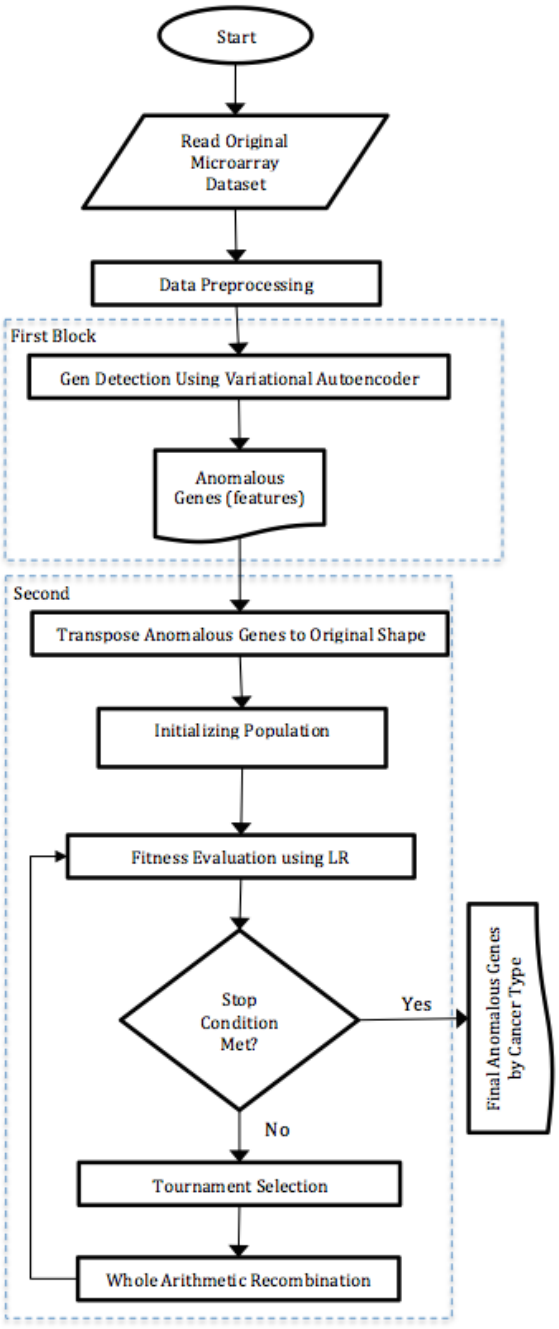

Fig. 1. The workflow of our proposed method

4.3.1. Chromosome Representation

Regarding to the problem of feature selection, we consider each chromosome as an array of float expression values with maximum length of 300 genes. It might be less than this number but definitly no more. In fact the length of each chromoseome equals to the number of outliers found in the first phase.

4.3.2. Population

We set the population size to 50. Inspite of the standard genetic algorithm, which the first population is generated randomly, in this adjusted genetic algorithm (AGA), we are benefiting from choosing the total population from the VAE outtput selectively for the sake of exploitation meaning by the anomalous genes found in the previous step.

4.3.3. Fitness Function

In order to evaluate the fitness of chromosomes, we have employed Logistic Regression (LR) and used classification accuracy to distinguish if the generated offsprings are acceptable or not. We use liblinear as a solver in LR setting. We have used voting technique consisting of SVM, MLP, Naïve Bayes, Random forest, and LR, but for our problem, LR achieves better generalization than the other mentioned methods.

4.3.4. Operators

From the second generation onwards, chromosomes are selected based on fitness. The selection operator used in our solution is tournament with size 3. This technique is effective in maintaining diversity while focusing on promising solutions. Crossover probability and mutation probability are set to 0.5 and 0.2 respectively.

5. EXPERIMENTAL RESULTS

In this section, we describe the performance of the both phases to select effective genes in disease occurrence on the Leukemia dataset having 72 samples and 7129 genes, and DLBCL dataset with the shape of 77 samples and 7071 genes as features. Our proposed method is implemented using the Python programming language in Google Colab environment. Also, the experimental results obtained on a T4 GPU. We have used Logistic Regression Algorithm to evaluate the selection of final genes. In fact, we have implemented the fitness function of the Genetic model with LR algorithm and we have reported the classification accuracy based on the obtained fitness value, which is actually the classification accuracy. In this implementation, we have used the 5Fold technique to perform each validation with the aim of reducing the risk of overfitting.

Table 1 and 2 show comparisons of classification before and after feature selection regard to both datasets. We have reported this data in order to show the effect and application of feature selection in increasing the accuracy of classification. Therefore, we conclude that reducing the number of effective genes improves the classification accuracy. Reducing overfitting is one of the important advantages of feature selection. Indeed, the smaller the overfitting, the better the model will be. Especially in cases where the number of features is large, the model learns well on train data, but it will not have the same power to generalize on test data.

Due to the stochastic nature of the genetic algorithm and the deep network, the results obtained in several runs have been recorded. We have announced the results based on the highest accuracy obtained from several independent runs on the dataset. The accuracy of the logistic classification after five executions along with the number of anomalous features, based on which we have presented the classification done related to each execution in Table 3.

The difference between the worst and the best accuracy scores obtained is less than or equals to 13, which indicates that with the minimum number of features having more entropy, better accuracy can be achieved.

We have also plotted the accuracy score on the training and validation data. As shown in Figure 2, we achieved an accuracy of 1.0 on the training data as a result of the classification process with the LR algorithm. But the important point is the growing trend of the validation score with the increase in the number of training samples.

To show how VAE works in our proposed method, and actually feature selection based on the distribution function, we have shown the distributions of two non-effective features that are not selected (a) and two effective features that are selected (b) in Figure 3. These plots show that the use of VAE in the selection of genes can be very effective. In addition, as a clue, which of course requires biological knowledge to validate, we have reported the names of 95 genes effective in leukemia 129 genes in Table 4.

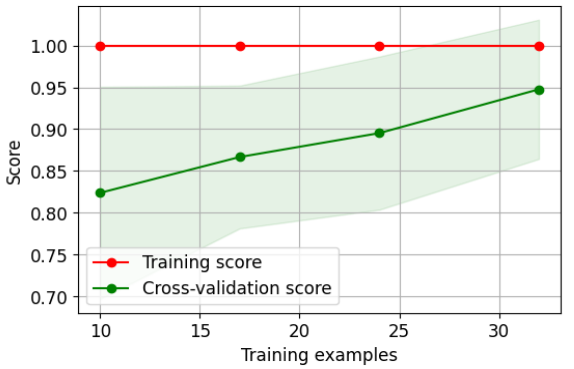

Fig. 2. Training and cross-validation score using LR with respect to the sample count

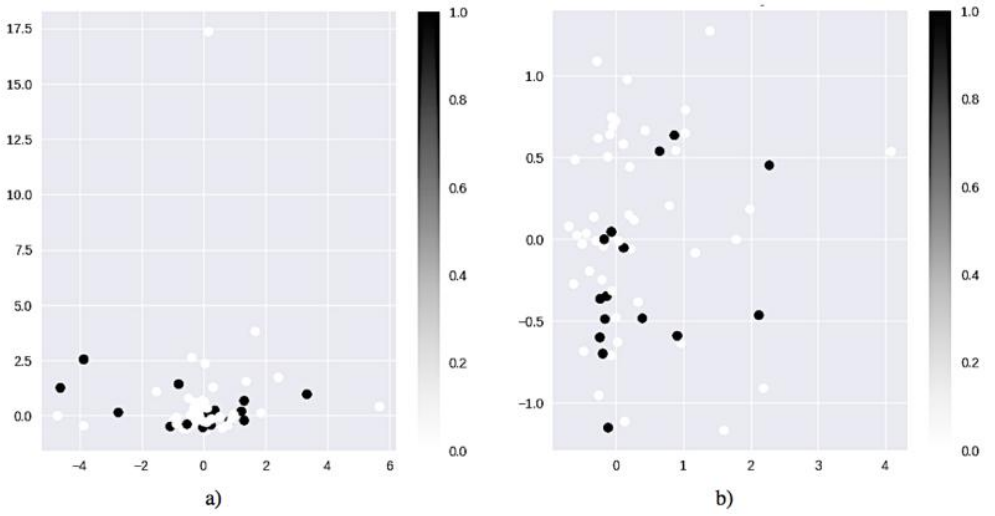

Fig. 3. Distribution of two features in the presence and absence of VAE.

Table 1. Classification comparisons before and feature selection on Leukemia dataset

Dataset1				
Test main data logistic				
	Precision	Recall	F1-score	Support
0	0.90	0.90	0.90	20
1	0.86	0.86	0.86	14
Accuracy			0.88	34
Macro Avg	0.88	0.88	0.88	34
Weighted Avg	0.88	0.88	0.88	34
Test feature selected genetic logistic				
	Precision	Recall	F1-score	support
0	0.95	0.95	0.95	20
1	0.93	0.93	0.93	14
Accuracy			0.94	34
Macro Avg	0.94	0.94	0.94	34
Weighted Avg	0.94	0.94	0.94	34

Table 2. Classification comparisons before and after feature selection on DLBCL dataset

Dataset 2

----------Test main data logistic----------

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	12
1	1.00	1.00	1.00	4
Accuracy			1.00	16
Macro Avg	1.00	1.00	1.00	16
Weighted Avg	1.00	1.00	1.00	16

----------Test feature selected genetic logistic----------

	Precision	Recall	F1-score	Support
0	1.00	0.92	0.96	12
1	0.80	1.00	0.89	4
Accuracy			0.94	16
Macro Avg	0.90	0.96	0.92	16
Weighted Avg	0.95	0.94	0.94	16

Table 3. Number of selected genes after a VAE execution and 5 executions of GA with the accuracy scores of each GA run.

Dataset	Index	Initial state	VAE	GA Runs				
				1st	2nd	3rd	4th	5th
Leukemia	Feature Count	7129	3564	58	43	127	66	61
	Accuracy	–	–	%85	%82	%94	%85	%94
DLBCL	Feature Count	7071	3535	244	206	129	139	144
	Accuracy	–	–	%81	%82	%94	%94	%94

Table 4. Comparison of our method with some other methods on Leukemia dataset

Method	Method name	Accuracy
Hybrid (Filter & genetic)	RefFPSO [37]	90%
Filter	TAaci [38]	96.25%
Wrapper	Ivsm-rfe [39]	94%
Our Method	DLAGA	94%

Furthermore, the top ten genes, which expressed abnormally in AML or ALL patients of the Leukemia dataset, have been revealed in figure 4. As already mentioned, having accurate knowledge of which genes play key role in any type of cancer is very promising in the treatment of cancer. Table 5 shows a comparison of the proposed method with other feature selection methods on the same Leukemia gene expression data.

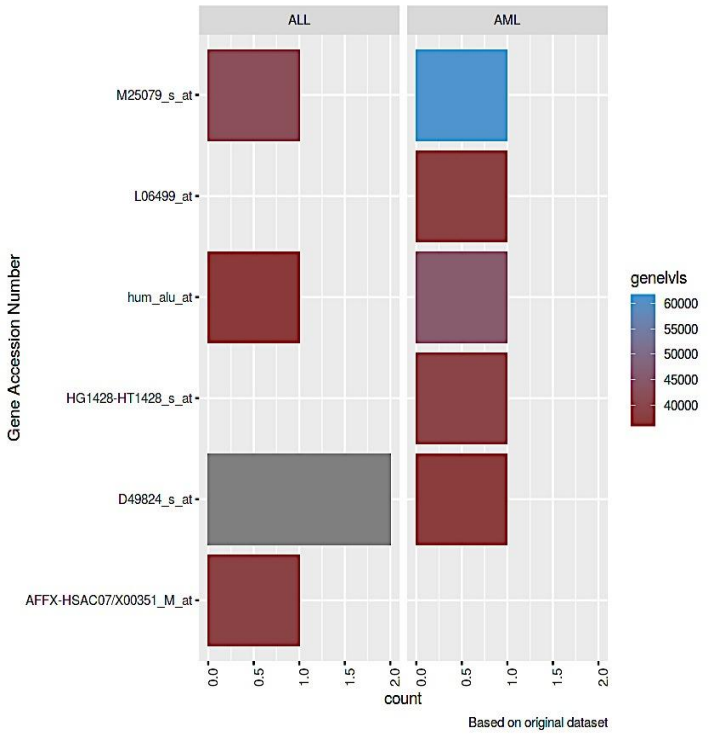

Fig. 4. Top 10 genes expresses in AML and ALL patients' genes of Leukemia dataset.

As it can be seen, we achieve a reasonable accuracy score. However, in some other articles, authors could achieve better scores. In the issue of selecting genes from microarray datasets, it is important to note that the proposed method should receive an acceptable score from several aspects. In order to discuss the selection of genes, apart from the criterion of classification accuracy, execution speed, complexity of time and space, as well as maintaining the relationship and meaning between genes are very important. Although our method has not reached the most in terms of accuracy score, it has worked well in terms of efficiency from the gene selection point of view while maintaining optimality

We believe that the correct and proportionate use of defined methods in pre-processing plays a decisive role in the success of any method. In our method -DLAGA- we paid special attention to this point so that it is logical and appropriate to the type of data, and does not conflict with the purpose of selecting anomalous genes.

6. CONCLUSION AND DISCUSSION

For the problem of selecting genes in microarray type data that face the challenge of high dimensions, our suggestion is to do the work in two consecutive phases and through the non-normal distribution function of genes.

In the pre-processing part, which is one of the most important steps, we started by transposing the original data in order to change the position of features and samples with each other. Then, by feeding the transposed data to the deep network (VAE), we found the distribution of each gene. Then we identified the genes that are out of the normal distribution as anomaly features that are more effective in the occurrence of diseases. In the pre-processing stage, in addition to transposing the data, we have used the RobustScaler method to standardize the data, which is a well-known normalization method in situations where there are outliers in the dataset. By performing this technique, a significant improvement on the results can be shown. To the best of our knowlege, the distribution function is very efficient to identify anomalous data. Therefore, in the first phase's task is set to reduce the dimensions of genes. With the help of a VAE deep network, the expression distribution of each gene was found and those that did not meet the normal distribution level were identified as abnormal genes and were selected for the next phase.

In the second phase, we transpose the abnormal genes selected from the previous phase back to the initial state. Now the data is ready to be fed as input to the genetic algorithm for more optimal selection of features. As a result, further optimal dimension reduction will be done at this phase on anomalous features. Due to the large number of

genes in the microarray data, performing any analysis is expensive and time-consuming. On the other hand, performing classification and other operations on these data requires the existence of real and intact genes. As a result, the use of feature extraction methods that change the main genes will not be useful. To solve these problems and to reduce the dimensions, performing these two steps is efficient and causes a significant increase in the classification accuracy. In the future works, we can apply a change in the fitness function of the genetic algorithm, so that instead of using machine learning algorithms, efficient neural network methods are used to increase the quality and accuracy. It is inevitable to consider the fact that human genomes change over several generations, however most of the studies conducted in the field of gene analysis have used outdated datasets or those that are not benchmarked, which causes unreliable or biased responses. Therefore, as a step forward, we can apply our method on the newly released microarray and RNA-Seq dataset. Also, with the help of distribution function obtained from different genes, we can model genes through neural network.

Data and code availability

All datasets are available at GEO database [40]. However, the data together with the source code are available at our GitHub repository [41].

Declaration

We acknowledge that we used ChatGPT to enhance the academic writing of our manuscript while ensuring the originality and integrity of our work.

Transparency Statement

The data supporting this study are available upon reasonable request to the corresponding author, subject to ethical and confidentiality considerations.

Acknowledgments

We would like to express our gratitude to all individuals who contributed to this project.

Declaration of Interest

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

REFERENCES

[1] Al Shanbari, N., Alharthi, A., Bakry, S. M., Alzahrani, M., Alhijjy, M. M., Mirza, H. A., Almutairi, M., & Ekram, S. N. (2023). Knowledge of cancer genetics and the importance of genetic testing: A public health study. Cureus, 15(8), e43016. https://doi.org/10.7759/cureus.43016

[2] Reda, B., Contardo, L., Prenassi, M., Guerra, E., Derchi, G., & Marceglia, S. (2023). Artificial intelligence to support early diagnosis of temporomandibular disorders: A preliminary case study. Journal of Oral Rehabilitation, 50(1), 31–38. https://doi.org/10.1111/joor.13383

[3] Salvadores, M., & Supek, F. (2024). Cell cycle gene alterations associate with a redistribution of mutation risk across chromosomal domains in human cancers. Nature Cancer, 1–17.

[4] Waarts, M. R., Stonestrom, A. J., Park, Y. C., & Levine, R. L. (2022). Targeting mutations in cancer. The Journal of Clinical Investigation, 132(8), e154943. https://doi.org/10.1172/JCI154943

[5] Grisci, B. I., Feltes, B. C., de Faria Poloni, J., Narloch, P. H., & Dorn, M. (n.d.). The use of gene expression

datasets in feature selection research: 20 years of inherent bias? Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, e1523.

[6] Nematzadeh, H., García-Nieto, J., Aldana-Montes, J. F., & Navas-Delgado, I. (2024). Pattern recognition frequency-based feature selection with multi-objective discrete evolution strategy for high dimensional medical datasets. Expert Systems with Applications, 123521.

[7] Potharlanka, J. L. (2024). Feature importance feedback with deep Q process in ensemble-based metaheuristic feature selection algorithms. Scientific Reports, 14(1), 2923.

[8] Zhou, H., Wang, X., & Zhang, Y. (2024). Feature selection based on weighted conditional mutual information. Applied Computing and Informatics, 20(1–2), 55–68.

[9] Ali, W., & Saeed, F. (2023). Hybrid filter and genetic algorithm-based feature selection for improving cancer classification in high-dimensional microarray data. Processes, 11(2), 562.

[10] Zhao, T., Zheng, Y., & Wu, Z. (2023). Feature selection-based machine learning modeling for distributed model predictive control of nonlinear processes. Computers & Chemical Engineering, 169, 108074.

[11] Hastie, T., Tibshirani, R., & Friedman, J. (2009). Model selection and regularization. In The elements of statistical learning: Data mining, inference, and prediction (2nd ed., pp. 219–259). Springer.

[12] Xu, C., & Zhang, S. (2024). A genetic algorithm-based sequential instance selection framework for ensemble learning. Expert Systems with Applications, 236, 121269.

[13] Janneh, L. L., Zhang, Y., Hydara, M., & Cui, Z. (2023). Deep learning-based hybrid feature selection for the semantic segmentation of crops and weeds. ICT Express. https://doi.org/10.1016/j.icte.2023.07.008

[14] Wang, Z., Pei, C., Ma, M., Wang, X., Li, Z., Pei, D., ... & Xie, G. (2024). Revisiting VAE for unsupervised time series anomaly detection: A frequency perspective. arXiv preprint arXiv:2402.02820.

[15] Radovic, M., Ghalwash, M., Filipovic, N., & Obradovic, Z. (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. BMC Bioinformatics, 18(1), 1–14.

[16] Bouazza, S. H., Auhmani, K., Zeroual, A., & Hamdi, N. (2018). Selecting significant marker genes from microarray data by filter approach for cancer diagnosis. Procedia Computer Science, 127, 300–309. https://doi.org/10.1016/j.procs.2018.01.126

[17] Masoudi-Sobhanzadeh, Y., Motieghader, H., Omidi, Y., & Masoudi-Nejad, A. (2021). A machine learning method based on the genetic and world competitive contests algorithms for selecting genes or features in biological applications. Scientific Reports, 11(1). https://doi.org/10.1038/s41598-021-82796-y

[18] Ghosh, M., Adhikary, S., Ghosh, K. K., Sardar, A., Begum, S., & Sarkar, R. (2018). Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods. Medical & Biological Engineering & Computing, 57(1), 159–176. https://doi.org/10.1007/s11517-018-1874-4

[19] Ghosh, M., Begum, S., Sarkar, R., Chakraborty, D., & Maulik, U. (2019). Recursive memetic algorithm for gene selection in microarray data. Expert Systems with Applications, 116, 172–185. https://doi.org/10.1016/j.eswa.2018.06.057

[20] Seyyedabbasi, A. (2023). Binary sand cat swarm optimization algorithm for wrapper feature selection on biological data. Biomimetics, 8(3), 310.

[21] Guo, J., Jin, M., Chen, Y., & Liu, J. (2020). An embedded gene selection method using knockoffs optimizing neural network. BMC Bioinformatics, 21(1), 414. https://doi.org/10.1186/s12859-020-03717-w

[22] Sahu, B., & Dash, S. (2024). Optimal feature selection from high-dimensional microarray dataset employing hybrid IG-Jaya model. Current Materials Science, 17(1), 21–43.

[23] Yaqoob, A., Verma, N. K., & Aziz, R. M. (2024). Optimizing gene selection and cancer classification with hybrid sine cosine and cuckoo search algorithm. Journal of Medical Systems, 48(1), 10. https://doi.org/10.1007/s10916-023-02031-1

[24] Babichev, S., Liakh, I., & Kalinina, I. (2024). Applying the deep learning techniques to solve classification tasks using gene expression data. IEEE Access, 12, 28437–28448. https://doi.org/10.1109/ACCESS.2024.3368070

[25] Uzma, Al-Obeidat, F., Tubaishat, A., Shah, B., & Halim, Z. (2022). Gene encoder: A feature selection technique through unsupervised deep learning-based clustering for large gene expression data. Neural Computing and Applications, 34(11), 8309–8331. https://doi.org/10.1007/s00521-020-05101-4

[26] Akhavan, M., & Hasheminejad, S. M. H. (2023). A two-phase gene selection method using anomaly detection and genetic algorithm for microarray data. Knowledge-Based Systems, 262, 110249.

[27] Xie, J., Rao, J., Xie, J., Zhao, H., & Yang, Y. (2024). Predicting disease-gene associations through self-supervised mutual infomax graph convolution network. Computers in Biology and Medicine, 108048. https://doi.org/10.1016/j.compbiomed.2024.108048

[28] Mai, S., Zheng, S., Yang, Y., & Hu, H. (2021). Communicative message passing for inductive relation reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 5, pp. 4294–4302).

[29] Xuan, P., Meng, X., Gao, L., Zhang, T., & Nakaguchi, T. (2022). Heterogeneous multi-scale neighbor topologies enhanced drug–disease association prediction. Briefings in Bioinformatics, 23(3), bbac123.

[30] Peng, Z., Huang, W., Luo, M., Zheng, Q., Rong, Y., Xu, T., & Huang, J. (2020, April). Graph representation learning via graphical mutual information maximization. In Proceedings of the Web Conference 2020 (pp. 259–270).

[31] Ino, K., Utagawa, Y., & Shiku, H. (2023). Microarray-based electrochemical biosensing.

[32] Gouda, W., Tahir, S., Alanazi, S., Almufareh, M., & Alwakid, G. (2022). Unsupervised outlier detection in IoT using deep VAE. Sensors, 22(17), 6617. https://doi.org/10.3390/s22176617

[33] Cai, Z., Yang, X., Zhou, M. C., Zhan, Z. H., & Gao, S. (2023). Toward explicit control between exploration and exploitation in evolutionary algorithms: A case study of differential evolution. Information Sciences, 649, 119656. https://doi.org/10.1016/j.ins.2023.119656

[34] Wang, A., Liu, H., Yang, J., & Chen, G. (2022). Ensemble feature selection for stable biomarker identification and cancer classification from microarray expression data. Computers in Biology and Medicine, 142, 105208.

[35] Jin, Z., Huang, Z., Wu, C., Zhang, F., Gao, Y., Guo, S., ... & Wu, J. (2024). Molecular insights into gastric cancer: The impact of TGFBR2 and hsa-mir-107 revealed by microarray sequencing and bioinformatics. Computers in Biology and Medicine, 108221.

[36] Miwa, D., Shiraishi, T., Duy, V. N. L., Katsuoka, T., & Takeuchi, I. (2024). Statistical test for anomaly detections by variational auto-encoders. arXiv preprint arXiv:2402.03724.

[37] Liu, M., Xu, L., Yi, J., & Huang, J. (2018). A feature gene selection method based on ReliefF and PSO. https://doi.org/10.1109/ICMTMA.2018.00079

[38] Taşci, A., İnce, T., & Güzelış, C. (2017). A comparison of feature selection algorithms for cancer classification

through gene expression data: Leukemia case.

[39] Kr, K., Kv, A. R., & Pillai, A. (2019). An improved feature selection and classification of gene expression profile using SVM (Vol. 1). https://doi.org/10.1109/ICICICT46008.2019.8993358

[40] National Center for Biotechnology Information. (n.d.). GEO DataSets. https://www.ncbi.nlm.nih.gov/gds/

[41] Yassiaap. (2024). Yassiaap/DLAGA [GitHub repository]. GitHub. https://github.com/Yassiaap/DLAGA.git