Volume 9, Issue 3, Article 18 (2025)

Signal Processing and Renewable Energy (SPRE)

https://doi.org/10.57647/j.spre.2025.090318

Deep Learning Based Cancer Classification Using Selected Gene Expression Data

Setare Tabasi[1], Iman Ahanian [2,*], Nasrin Amiri[2]
Nader Jafarnia Dabanloo[3], Hamide Barghamadi[1]

[1] *Department of Biomedical Engineering, ST.C., Islamic Azad University, Tehran, Iran*

[2] *Department of Electrical Engineering, ST.C., Islamic Azad University, Tehran, Iran*

[3] *Department of Biomedical Engineering, SR.C., Islamic Azad University, Tehran, Iran*

***Corresponding author:**. ahanian@iau.ac.ir

Original Research

Received:
24 June 2025

Revised:
19 July 2025

Accepted:
4 August 2025

Published in Issue:
30 September 2025

Abstract

Cancer incidence has yearly surged by 23.6 million in 2019, with 10 million deaths from cancer reported in the same year, signifying a 26% and 21% upsurge in cases and deaths over the last decade. Gene expression (GE) analysis is a robust technique for the early diagnosis and classification of cancers by identifying unique molecular characteristics in diverse organs. Unlike other methods, identifying cancer-related genes from GE profiles results in more effective and tailored therapies. This technique, however, suffers from shortcomings such as huge amounts of GE data available, making data extraction and analysis a challenging task, a dearth of inclusive databases, and poor access to GE data extraction technology in some regions. This research proposes a deep learning (DL)-based minimum-redundancy-maximum-relevance (mRMR) technique for the precise identification of six cancer types using GE data, where DL-mRMR merges DL and feature selection (FS) to pick effective features and reduce dataset dimensions. Entrenched in FS, DL-mRMR downsizes the volume of input features reflecting similarities between cancers and within samples of a given tumor, and eventually balances the samples in the database, thereby diagnosing cancer types using a DL classifier. TCGA database simulations reveal a 99% accuracy (ACC) of DL-mRMR in differentiating breast, colorectal, kidney, and lung cancers based on GE data. In addition, due to the reduction in the size of the input sample vector, the complexity of the classifier algorithm and the number of samples required for its training are lower than similar approaches.

Keywords: Gene expression, Cancer classification, Deep learning, Feature reduction, mRMR

Cite this article: Tabasi S., Ahanian I., Jafarnia Dabanloo N., Barghamadi H , Deep Learning Based Cancer Classification Using Selected Gene Expression Data, *Signal Process. Renew. Energy.* 9 **(**3) Article 18 (2025). https://doi.org/10.57647/j.spre.2025.090318

1. Introduction

Cancer refers to a group of diseases emanating from uncontrolled and inadvertent growth and proliferation of cells in the body. Cancer is commonly characterized by interrupted regulatory mechanisms of normal growth, proliferation, and cell death. Cancer cells are capable of invading or spreading to nearby tissues while possessing the ability to metastasize to other body organs. Overall, cancer arises from diverse environmental and genetic contributors and is associated with numerous physical symptoms. In addition to symptoms in the original location (i.e., primary tumors), cancer cells metastasize to other organs via spreading through the vasculature and lymphatic system by adopting mechanisms of destruction, invasion, and occupation. By 2024, the number of new cancer cases and the resulting cancer deaths were predicted to rise to 29.9 and 15.3 million, respectively [1]. Of all the

10 million new cancer cases diagnosed annually, one-third are preventable, and one-third are treatable upon early diagnosis**.**

According to epidemiological research, cancer is the third leading cause of mortalities in Iran after cardiovascular diseases and crashes [2], thereby requiring prompt actions on cancer control based on early prevention and screening and adopting versatile systems to collect, record, assess, and analyze information on new cancer cases. Importantly, cancer will progressively be a key contributor to the global burden of disease shortly. Statistically, of all 10 million new cancer cases yearly, 4.7 and 5.5 million are expected to arise in developed and less-developed nations, respectively. Regardless of cancer being overwhelming in developed nations, over half of all cases happen in developing nations. Indeed, cancer is the second cause of mortalities and disabilities in developed societies, and the third cause of death in developing societies [3].

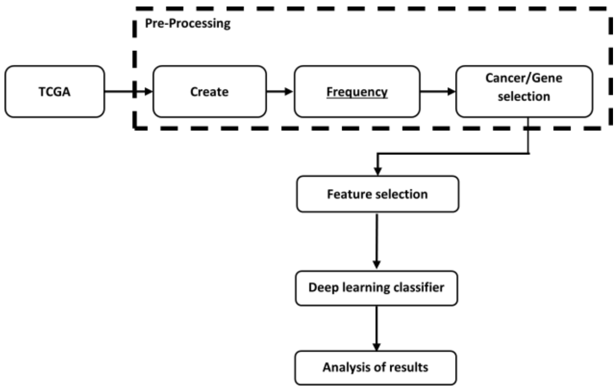

Figure 1. The proposed method

There is no single test for the precise diagnosis of cancer, and current procedures require a merge of physical examinations and tests, assessments of patient background, and diagnostic kits. Most cancers share mutual symptoms, which prompts researchers to develop novel tools for the early and accurate diagnosis and differentiation of cancer. Presently, diagnostic techniques include imaging, tumor markers, biopsy, endoscopic observations, surgery, and genetic assessments. Genetically, a novel technique identifies genetic biomarkers for each cancer by studying GE profiles. It is, however, a daunting task to identify genes that contribute to each cancer, as human cells encompass plentiful genes that each inconsistently affect cancer incidence or experience variations in their specific characteristics upon cancer occurrence. Though DL and cognitive models have recently fostered cancer prediction and diagnosis [4-6], some unresolved issues remain when using GE profiles for this purpose due to the following factors:

- Multifaceted gene structures
- Impacts from exterior factors, cancers, and diseases of the expression of each gene, leading to wrong diagnosis
- Limited access to public GE databases and insufficient DL techniques available
- The number of features exceeds that of samples due to massive counts of genes

This research proposes the DL and FS-based mRMR technique to differentiate breast, colorectal, kidney, and lung cancers using GE data. As such, this research contributes to the literature by 1) proposing a hierarchical approach with distinct domains (DL and FS) for cancer classification, 2) proposing the mRMR technique to specify effective features on diverse GE profiles in various cancers, and 3) using binary DL classifiers to detect cancer types.

2. Literature review

When processing microarray data obtained from GE profiles, the sought-after FS techniques include embedded, wrapper, filter, and ensemble procedures [7]. Embedded techniques directly incorporate FS into the learning process of the classification model. Wrapper techniques employ a classification model as part of the FS process. Filtering techniques (e.g., mRMR) assess and select features independent of the classification model and, instead, based on statistical criteria or rankings. Ensemble techniques employ and merge multiple FS methods for enhanced precision.

Kang et al. proposed rLGenSVM (relaxed Lasso and generalized multi-class support vector machine – SVM) as a novel embedded classification model for tumor identification to reduce the dimensionality of GE data. As reported, rLGenSVM optimizes (by reducing) the number of feature genes and delivers significantly higher classification ACC [8]. Zhu et al. developed an embedded method for FS using kernel parameters descent SVM (KPD-SVM). They employed a gradient descent algorithm to search for the best subset of features [9]. Elsewhere, Zhu et al. proposed KP-CSSVM as an FS and classification method with SVM for cases when the classes are distributed asymmetrically. Mishra et al. [10] developed SVM-BT-RFE as an improved gene selection framework that merges SVM-T-RFE (Recursive Feature Elimination t-test SVM) and SVM-RFE (Recursive Feature Elimination SVM). Zhang et al. [11] proposed the embedded FS method “MPMCSVM-RFE” (a multiple version of SVM-RFE) for FS and multiclass classification [12]. Besides, similar approaches are available elsewhere (see, e.g., [12-16]).

Particle swarm optimization (PSO) is a popular metaheuristic algorithm holding widespread applications as a wrapper method for FS in microarray data processing. Jain et al. reported an improved version of the binary PSO algorithm (i.e., the iBPSO wrapper) for enhancing classification ACC [17].

Moradi et al. reported HPSO-LS as a new wrapper method for FS based on PSO, which works by merging the PSO algorithm and a local search strategy. This wrapper employs a local search algorithm to direct the search process in selecting genes based on their correlation [18]. Garibay et al. proposed a wrapper method as a modified version of the inertial geometric PSO algorithm for selecting genes in DNA microarray datasets. They worked with diverse classifiers, including SVM, self-organizing map (SOM), back propagation neural network (BPNN), Naïve Bayes (NB), decision trees (CART), artificial

10.57647/j.spre.2025.0903.18

immune recognition system (AIRS), and PSO with C4.5 decision tree (PSODT) [19].

Mohapatra et al. reported a modified PSO-based wrapper method with SVM, NB, and k-nearest neighbor (KNN) classifiers for FS from microarray datasets of prostate cancer, leukemia, and colon tumors [20]. Chen et al. used a PSO-based decision tree classifier for gene selection in ten cancer datasets. They first worked with PSO to generate a subset of genes, followed by assessing the ACC of the generated subset using the C4.5 classifier [21].

The firefly algorithm (FA) [22] has been similarly used in the literature. Almugren et al. proposed the FA-based wrapper FS method "FF-SVM" and evaluated its ACC by using five standard microarray datasets and SVM classifiers. Jinthanasatian et al. [23] proposed a novel approach for cancer detection that intelligently merges a neuro-fuzzy system and FA. As reported, the FA used in this approach optimizes the system parameters and concurrently selects and classifies effective features [23]. Raguntthar et al. developed a wrapper-based FS method by merging the artificial bee colony (ABC) algorithm with stochastic diffusion search (SDS). As reported, the hybrid ABC-SDS method outperforms both SDS or ABC algorithms alone in delivering improved performance [24]. Pratiwi et al. used the genetic bee colony (GBC) algorithm for FS and the parameter "accuracy" to evaluate the performance of their proposed method "MBP-CGP (Conjugate Gradient Backpropagation with Modified Polak Ribiere), which was used as a classifier. By selecting features of up to 47 to 51% of all datasets, the ACC of results using MBP-CGP and GBC classifications spanned from 88.75 to 100% [25]. Tawhid et al. developed an FS method based on the binary whale optimization algorithm (BWOA). BWOA performance was evaluated using three classifiers and 32 datasets from the UCI machine learning (ML) repository [26].

Zakari et al. proposed the FS method "GOFS" using a modified version of the grasshopper optimization algorithm (GOA). For performance enhancement, they modified the GOA to balance the broad search (exploration) and the focused search (exploitation). Such an improvement was achieved using a mathematical model to simulate attractive and repulsive forces between grasshopper individuals within a swarm [27]. Chatra et al. developed an FS method based on the binary bat algorithm using an innovative fitness function. As reported, this function reduces and increases intra-class and inter-class distances, respectively, to enhance classification ACC [28]. Ghosh et al. proposed a recursive memetic algorithm (RMA) as a wrapper-based FS method [29]. Similarly, Allam et al. proposed the binary teaching learning-based optimization (FS-BTLBO) algorithm as a new wrapper-based FS method [30]. Sharma et al. developed a wrapper-based FS method based on improved regularized linear discriminant analysis for cancer classification [31]. Other wrapper-based approaches are similarly reported in the literature (see, e.g., [32-36]).

Sharbaf et al. employed the Fisher criterion as a filter to reduce the number of initial genes and computational complexity [37]. Jain et al. proposed a hybrid method that merges Correlation-based FS (CFS; as a filter) with improved PSO (iBPSO; as a wrapper) [38].

Chinnaswamy et al. developed a hybrid method for FS that merges the correlation coefficient (as a filter) and the PSO algorithm (as a wrapper) [39]. Pashaei et al. proposed a hybrid method encompassing a Binary Black Hole Algorithm (BBHA; as a wrapper) and Random Forest Ranking (RFR; as a filter) [40]. Elsewhere, Dabba et al. [41] proposed a hybrid FS method using mRMR (as a wrapper) based on hybridization between quantum computation and moth flame optimization (MFO), with SVM serving as the classifier [41]. Baliarsingh et al. employed ANOVA (as a statistical filter) and hybridization between the enhanced Jaya (EJaya) algorithm and forest optimization algorithm (FOA) as a wrapper [42]. Alanni et al. merged Gain Ratio (GR; as a filter) and improved GE programming (IGEP; as a wrapper) algorithms for FS [43]. The same works on filter and hybrid approaches are presented elsewhere in the literature (see, e.g., [44-48]).

As with the literature review, the leading challenges in FS based on GE data include the following:

• Data scattering and uncorrelation in different patients with mutual cancer and massive counts of existing genes
• Common features of different cancers and external factors affecting GE profiles
• The possibility of multiple cancers occurring simultaneously and causing errors in the training process
• Overall, there is no single method to diagnose a set of cancers based on GE profiles. Similarly, there is a dearth of a segregated gene set for a specific cancer to serve as a reliable marker for diagnoses.

3. The proposed method

The proposed method operates with mRMR feature reduction and DL (convolutional neural network – CNN) (Figure 1). First, data (e.g., on GE profiles) obtained from the database are analyzed concerning frequency distributions and then entered into the mRMR block for feature reduction. The remaining genes (as inputs) enter the DL classifier to deliver final results. Overall, this method trains several different binary networks (one network for each cancer), where several types of cancer are assumed to be in the input data.

3.1. Database

The Cancer Genome Atlas (TCGA) was used as the comprehensive database [33], which is jointly led by the National Institutes of Health (NIH), the National Cancer Institute (NCI), and the National Human Genome Research Institute (NHGRI). TCGA fosters identifying all key genomic alterations in the main cancers and their subtypes. In TCGA, data have evolved with fast progress in genetic sequencing. Notably, the technologies adopted to measure specific variables (e.g., mRNA expression, methylation, or single-nucleotide polymorphisms [SNPs]) can be restructured over time. TCGA has applications in numerous cancer classification research (see, e.g., [34-38]).

Table 1. Number of samples available in the TCGA database for various cancers

Cancer	N of samples	Cancer	N of samples
Adrenal Gland.xlsx	290	Pancreas	445
Bile Duct	9	Cervix	808
Bladder	408	Skin	500
Bone Marrow and blood	1000	Soft tissue	134
Bone	98	Stomach	447
Brain	913	Testis	135
Breast	1000	Thymus	90
Esphagus	222	Thyroid	945
Eye	80	Uterus	879
Head and Neck	645	colorectal	808
Lymph Nodes	519	kidney	1000
Nervous System	7	liver	409
Ovary	497	lung	1000

For each cancer, GE data are identified based on clustering, with processing all the GE samples. Table 1 presents the count of samples available for various cancers.
To process all the samples available, the frequency distribution table for various genes is first extracted, followed by selecting more key genes and more frequent cancers and then using them in feature reduction and classification.

3.2. Feature Reduction

mRMR is broadly used to select GE data in cancer classification. For instance, Gulande et al. reported "mRMR merged with ML techniques" as a potent strategy to diagnose lung cancer based on GE data [49]. Similarly, Yaqoob merged mRMR with the Northern Goshawk Algorithm (NGHA) to classify breast cancer. Elsewhere, Khaing et al. merged mRMR with SVM (as a classifier) to reduce GE data to diagnose breast, renal, colon, lung, and prostate cancers [50].
As an FS technique, mRMR minimizes information overlap between features, while ensuring each selected feature can significantly predict the outcome (targetvariable) [51, 52]. mRMR markedly boosts the ACC and efficiency of the system, optimally balancing relevance (to the target) and redundancy (reduction of overlap). mRMR efficacy in high-dimensional data has been validated in some research [53].
In mRMR, relevance refers to the contribution (influence) of a feature to the prediction of the outcome variable, which is the presence or absence of cancer in this research. Relevance is measured using the index "mutual information", which quantifies the degree of information shared between a feature and the outcome. For a feature Xi and an output Y, relevance is calculated from the Eq. 1:

$$I(X_i;Y) = \sum_{x\in X_i}\sum_{y\in Y} p(x,y)\,log\left(\frac{P(x,y)}{p(x)p(y)}\right) \quad (1)$$

Where Xi is one of the GE data in the ith person, and Y is a binary output that clarifies the presence or absence of cancer or no cancer in that person. The function p(x,y) specifies the mutual probability of each feature Xi occurrence with Y, where these values are extracted from the training data. As such, p(x) and p(y) are marginal probabilities for Xi and Y, signifying the independent distribution of each in the data set. Eq. 1 uses a base-2 logarithm, where its output is in bits, and quantifies a reduction in the uncertainty about Y when knowing the value of Xi.
Since GE values are continuous, p(x) and p(x,y) cannot be directly calculated. Thus, the features become discrete in most cases before calculating mutual information. Here, discretization, which is mostly achieved using methods such as binning, estimates the probability distribution for continuous variables. Bigger I(Xi;Y) values imply that the feature Xi is highly potent in predicting the output and, hence, is prioritized in FS by mRMR.
Redundancy assesses the volume of information overlap between a candidate feature (gene) and previously selected features (genes). Indeed, redundancy (as validated by recent research, e.g., [51]) quantifies redundant information provided by a feature compared to other available features. Overall, redundancy is defined as the mean volume of mutual information
between the candidate feature Xi and all other features (Xj):

$$I_{x_i} = \frac{1}{|S|^2}\sum_{X_i,X_j\in S,i\neq j} I\left(X_i;X_j\right) \quad (2)$$

Where S denotes the total number of remaining features, I(Xi;Xj) quantifies the volume of information shared between the candidate feature Xi and each feature Xj from the existing feature set (as obtained from Eq. 2). A higher I(Xi;Xj) value signifies a higher overlap of information between the features, with its mean value specifying the similarity of a feature with all the existing features. As such, mRMR minimizes this value to ensure that the final set of features is informationally diverse.
For FS, mRMR instantaneously integrates relevance and redundancy into a single objective function, where the function is obtained from the Eq. (3) [51]:

$$\phi(X_i) = I(X_i;Y) - I_{x_i} \quad (3)$$

Where ϕ(Xi) is the final score of Xi, with the first part [i.e., I(Xi;Y)] denoting the relevance or predictive power of feature Xi, and the second part quantifying redundancy.
Overall, features with higher ϕ(Xi) are more relevant.

3.3. Deep learning classifier

Fig.2 illustrates the structure of the DL classifier encompassing five different layers.

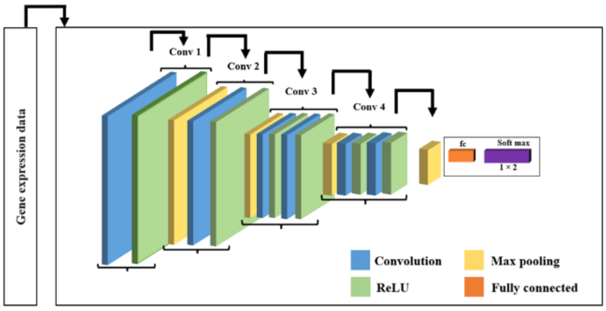

Figure 2. DL binary classifier.

Convolution layer
The convolutional layer is a primary component of CNN in DP that extracts key features from GE data. This layer applies a set of filters (kernels) with the convolution operator * (see Eq. 4) on the input vector to identify diverse input features.

$$x[n] * k[n] = \sum_{m=-\infty}^{\infty} x[m]x[n-m] \tag{4}$$

Where x[n] is the input and k[n] is the kernel.
ReLU
Typically, there is a nonlinear activation function after the convolution layer, with the rectified linear unit (ReLU) serving as the most interesting nonlinear activation function (Eq. 5).

$$\begin{aligned} ReLU(x) &= max(0,x) \\ &= \begin{cases} x & \text{if } x > 0, \\ 0 & x \le 0 \end{cases} \end{aligned} \tag{5}$$

Max pooling layer
Max pooling is a layer in CNN that reduces the dimensionality of the features built in the middle convolutional layers by diminishing the size of data (while preserving important information) and enhancing model efficiency. In this layer, the highest value from each section is reflected as the output.
Fully connected layer (FCL)
In this layer, neurons are connected to all neurons in the previous layer, where it first multiplies all the previous layer outputs in a weight matrix and then transfers them to the next layer. The last FCL merges the features for cancer classification. Hence, the last FCL output equals the number of classes, which is 2 in this research.
Softmax Layer
This layer normalizes the FCL's output, with the output of this layer having positive numbers summing up to one, where these numbers are reflected as probabilities of belonging to each class.

3.4. Evaluation criteria

The following evaluation criteria are used to evaluate the proposed method:
TP: Shows the correctly detected second cancer in data.
FP: Shows the number of wrongly detected second cancers.
FN: Shows the number of wrongly detected first cancers.
TN: Shows the correctly detected first cancer in data.
Precision (Pr)
This criterion specifies the ratio of correctly detected second cancer to the total data (encompassing the second cancer).

$$Pr = \frac{TP}{TP + FP} \tag{6}$$

Recall (Re)
This criterion denotes the proportion of correctly diagnosed second cancers to the total outputs diagnosed by the model as second cancers.

$$\text{Re} = \frac{TP}{TP + FN} \tag{7}$$

The F-score (F1)
Represents the ratio of correctly diagnosed second cancers to the total count of segments containing correctly diagnosed second cancers, divided by the total count of model errors.

$$\begin{aligned} F1 &= 2\frac{Re\,c\,all \times Precision}{Re\,c\,all + Precision} \\ &= \frac{TP}{TP + \frac{(FP+FN)}{2}} \end{aligned} \tag{8}$$

ACC
Denotes the ratio of all correctly diagnosed cancers to the total errors of the proposed method.

$$ACC = \frac{TP + TN}{FP + FN} \tag{9}$$

4.Results

4.1. Preprocessing

First, the frequency distribution of diverse genes that effectively contribute to cancer(s) is assessed after reading the information on all the cancers in the database. This section covers "gene selection" and "cancer selection". For cancer selection, the number of samples is considered for concurrent analysis and determining the significance of that cancer. Here, breast, colorectal, kidney, and lung cancers are chosen from Table 1, all holding a sufficient number of samples. Fig. 3 quantifies genes shared between diverse cancers. As depicted, 243 diverse genes are shared by three or more cancers, which are reduced to 29 genes shared among 20 cancers. Since this value is simply an initial criterion for gene selection, the GE data of 243 genes effectively contribute to at least three cancers and are reflected as input for feature reduction

4.2. Feature reduction and classification

After specifying genes that are mutual in more than three cancers during preprocessing, a binary classifier is trained separately to differentiate between two cancers, as disparities exist in common genes between diverse cancers. First, GE profiles for genes that are mutual between the two cancers (in the list of 243) are identified, followed by extracting the corresponding GE data. The obtained samples with their corresponding classes are then randomly split into two training (70% of samples) and test (30% of samples) groups to be used as input to the DL classifier post-FS. Notably, the count of training and test data is the same in each binary classifier when evaluating how the count of genes affects performance. Here, the count of genes in the classifier's input surges by five steps to assess the FS importance in differentiating cancers. Thus, variations in the classifier's efficiency upon raising the count of input GE data are assessed, thereby effectively illustrating how the classifier's output is affected by the count of genes employed. As such, the evaluation criteria are reported based on the test data, where six binary classifiers are trained. Table2 tabulates the count of genes selected before FS, the count of samples used in each classifier, and the specification of that classifier.

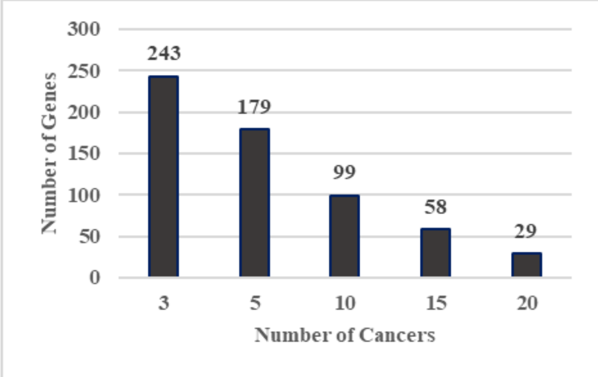

Figure 3. The count of genes shared between diverse cancers

Table 2. The binary classifiers trained in the second stage

No.	First class	Second class	No. of mutual genes before pre-processing	Total samples	Test samples	Training samples
1	Breast	Colorectal	52	1808	1266	542
2	Breast	Kidney	42	2000	1400	600
3	Breast	Lung	54	2000	1400	600
4	Colorectal	Kidney	52	1808	1266	542
5	Colorectal	Lung	57	1808	1266	542
6	Kidney	Lung	56	2000	1400	600

As supported by the results (Fig. 4), omitting excess genes from the input set is an effective strategy, with the classifier's best performance achieved with five selected genes. Conversely, a rise in the count of GE data reduces the classifier's performance. The ACC value exceeding 95% suggests the classifier's acceptable performance. Concerning similar performance in two classes, the best response is achieved from 20 genes.

As shown in Fig. 5, this classifier best performs (i.e., regarding proximity in the two classes and ACC) with 10 selected genes. Conversely, the classifier's performance decreases with the number of genes that surge, signifying the importance of feature reduction. As depicted in Fig. 6, this classifier best performs with 25 selected genes. Conversely, the classifier's performance reduces when the number of genes increases, signifying the importance of feature reduction.

As shown in Fig.7, the best performance for this classifier is obtained with 10 selected genes. However, the response pattern differs from previous scenarios under both higher or lower numbers of genes. As depicted in Fig.8, the patterns are similar, with all the mutual genes required for proper differentiation of the two classes. Though having five genes as inputs delivers the best ACC, other evaluation criteria deliver unacceptable results, implying the classifier's bias towards one of the two cancers in the output. As demonstrated in Fig. 9, the classifier's performance fluctuates markedly with the count of genes, though the performance exceeds 90% in most cases. Here, 35 genes deliver the best performance concerning the closeness of the response in the two classes. Table 3 compares the results in Fig.9 by representing the corresponding values.

As a summary, the proposed method is efficient in all cases, and its efficacy is markedly affected by the FS approach adopted in this research.

4.3. Comparison with previous research

Research has previously studied the TCGA database to classify cancer based on GE data while covering diverse cancers, genes, and networks. Mohammed et al. worked with a stacking ensemble DL approach to classify cancer based on TCGA data. They selected 173 genes after feature reduction and reported an ACC of 99.22% in differentiating breast, lung, colorectal, thyroid, and ovarian cancers [54]. Divate et al. employed a deep neural network (DNN) to classify 37 cancers using 13,250 genes from the

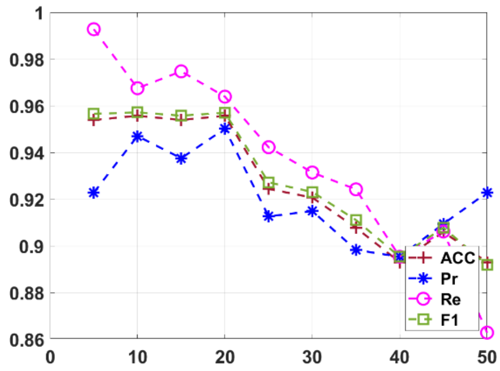

Figure 4. The count of genes shared between diverse cancers

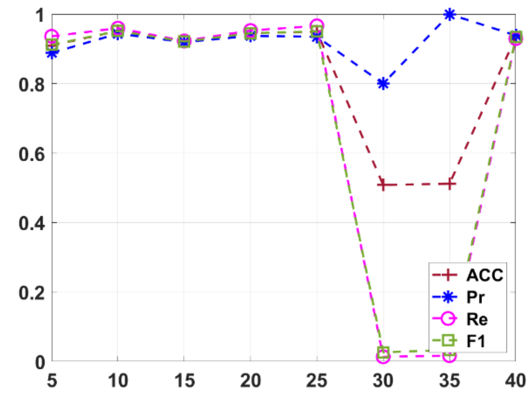

Figure 5. Evaluation criteria for the second classifier between Breast and Kidney cancers

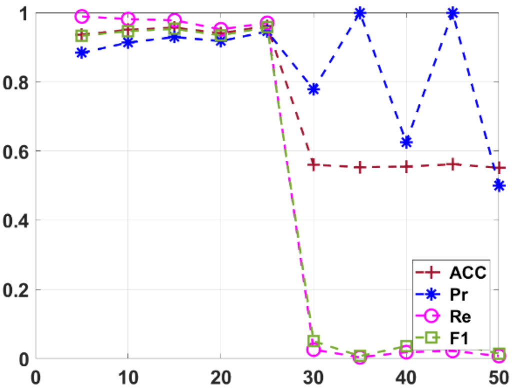

Figure 6. Evaluation criteria for the third classifier between Breast and Lung cancers.

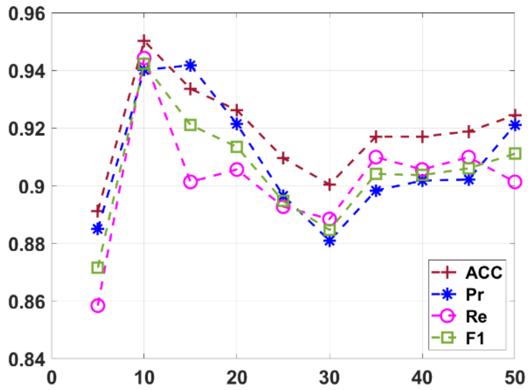

Figure 7. Evaluation criteria for the fourth classifier between Colorectal and Kidney cancers

Table 3. Evaluation criteria for the sixth classifier based on the count of genes

Number of genes	TN	FN	FN	TP	ACC	Pr	Re	F1
5	255	22	35	288	0.905	0.892	0.929	0.910
10	262	22	28	288	0.917	0.911	0.929	0.920
15	263	27	27	283	0.910	0.913	0.913	0.913
20	263	21	27	289	0.920	0.915	0.932	0.923
25	264	21	26	289	0.922	0.917	0.932	0.925
30	255	24	35	286	0.902	0.891	0.923	0.906
35	266	23	24	287	0.922	0.923	0.926	0.924
40	270	26	20	284	0.923	0.934	0.916	0.925
45	271	23	19	287	0.930	0.938	0.926	0.932
50	274	26	16	284	0.930	0.947	0.916	0.931
55	278	26	12	284	0.937	0.959	0.916	0.937

TCGA database, with the model achieving a predictive ACC of 97% at best [55]. Ramirez et al. worked with a Graph CNN to classify cancer based on the TCGA database. They considered 428 genes as input data, resulting in the classification of 33 cancers, with the final ACC spanning from 89.9 to 94.7% [56]. Notably, the number of genes in these studies is higher than that in the present research. By reporting overall accuracy, these studies have overlooked evaluating the evaluation entirely. Thus, the method proposed in the present research is acceptability efficient in all cases and the FS algorithm has markedly reduced the count of genes while enhancing classification efficiency.

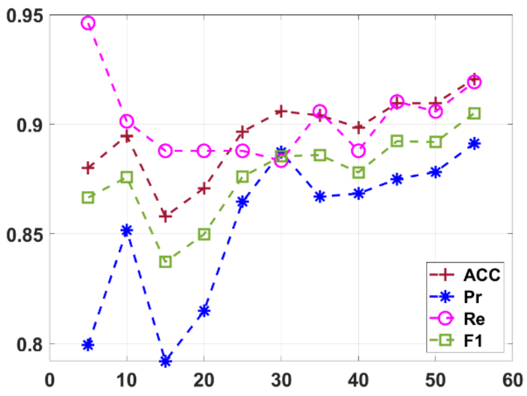

Figure 8. Evaluation criteria for the fifth classifier between Colorectal and Lung cancers

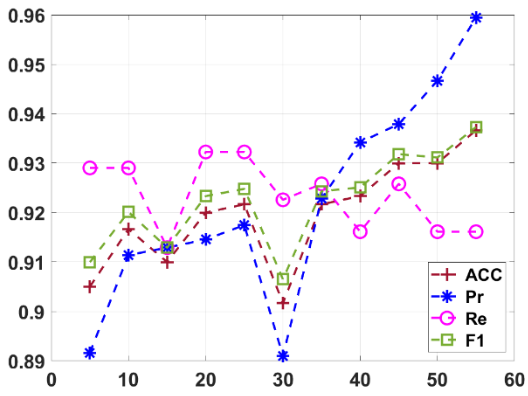

Figure 9. Evaluation criteria for the sixth classifier between Lung and Kidney cancers

5. Conclusion

GE data fosters cancer diagnosis by identifying abnormal patterns in cells. Analyzing GE data allows for a precise classification of cancers and ultimately delivers treatments tailored to each cancer. Considering GE data analysis challenges arising from the huge count of genes, this research proposes DL-mRMR to select genes that are more effective in cancer prediction and classification. At first, DL-mRMR runs an initial preprocessing by measuring the GE frequency table and reflecting the gene mutuality between diverse cancers. Next, it uses a system containing a DL-based classifier to select genes that are more effective in differentiating breast, colorectal, kidney, and lung cancers. Further, DL-mRMR thoroughly assesses diverse parameters, including precision, recall, F1, and ACC in all cases. Simulations on the TCGA database reveal an ACC of over 90% in all cases in differentiating breast/colorectal and kidney/lung cancers based on the expression profiles of genes that are mutual between cancer types. Due to reduced input sample vector volume, classification by DL-mRMR entails a less complex classifier with lower counts of samples required for its training. As such, while DL-mRMR holds at least three times fewer genes compared to the smallest gene set in other research, it delivers acceptable efficiency in cancer classification.
For future research, the efficiency and ACC of the proposed method to GE data errors can be considered. The impact of other genetic problems, such as unknown hereditary problems and genetic variations based on other diseases, on reducing the efficiency of the proposed system is another interesting topic. Furthermore, hardware resources for training the proposed approach in practical conditions (Due to the huge amount of GE data) can be considered in future research.

Authors Contribution
All the authors have participated sufficiently in the intellectual content, conception and design of this work or the analysis and interpretation of the data (when applicable), as well as the writing of the manuscript.

Availability of data and materials
The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Conflict of interests
The author states that there is no conflict of interest.

References

[1] Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2024 May;74(3):229-63. doi: https://doi.org/ 10.3322/caac.21834

[2] Seyedkanani E, Hosseinzadeh M, Mirghafourvand M, Sheikhnezhad L. Breast cancer screening patterns and associated factors in Iranian women over 40 years. Scientific Reports. 2024 Jul 3;14(1):15274.doi: https://doi.org/10.1038/s41598-024-66342-0

[3] Pramesh, C. S., Rajendra A. Badwe, Nirmala Bhoo-Pathy, Christopher M. Booth, Girish Chinnaswamy, Anna J. Dare, Victor Piana de Andrade, V.P., Hunter, D.J., Gopal, S., Gospodarowicz, M. and Gunasekera, S. "Priorities for cancer research in low-and middle-income countries: a global perspective." Nature medicine 28, no. 4 (2022): 649-657. doi: https://doi.org/10.1038/s41591-022-01738-x

[4] Manikandan G, Abirami S. Feature selection is important: state-of-the-art methods and application domains of feature selection on high-dimensional data. Applications in Ubiquitous Computing. 2021:177-96. doi: https://doi.org/10.1007/978-3-030-35280-6_9

[5] Alharbi, Fadi, and Aleksandar Vakanski. "Machine learning methods for cancer classification using gene expression data: A review." Bioengineering 10, no. 2 (2023): 173.doi: https://doi.org/10.3390/bioengineering10020173

[6] Liu, Hengrui, Zheng Guo, and Panpan Wang. "Genetic expression in cancer research: Challenges and complexity." Gene Reports (2024): 102042. doi:https://doi.org/10.1016/j.genrep.2024.102042.

[7] Alhenawi, Esra'A., Rizik Al-Sayyed, Amjad Hudaib, and Seyedali Mirjalili. "Feature selection methods on gene expression microarray data for cancer classification: A systematic review." omputers in biology and medicine 140 (2022): 105051. doi: https://doi.org/ 10.1016/j.compbiomed.2021.105051

[8] Kang, Chuanze, Yanhao Huo, Lihui Xin, Baoguang Tian, and Bin Yu. "Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine." Journal of theoretical biology 463 (2019): 77-91. doi: https://doi.org/ 10.1016/j.jtbi.2018.12.010

[9] Zhu, Haiqing, Ning Bi, Jun Tan, and Dongjie Fan. "An embedded method for feature selection using kernel parameter descent support vector machine." In Pattern Recognition and Computer Vision: First Chinese Conference, PRCV 2018, Guangzhou, China, November 23-26, 2018, Proceedings, Part III 1, pp. 351-362. Springer International Publishing, 2018. doi: https://doi.org/10.1007/978-3-030-03338-5_30

[10] Mishra, Shruti, and Debahuti Mishra. "SVM-BT-RFE: An improved gene selection framework using Bayesian T-test embedded in support vector machine (recursive feature elimination) algorithm." Karbala International Journal of Modern Science 1, no. 2 (2015): 86-96. doi: https://doi.org/10.1016/j.kijoms.2015.10.002

[11] Zhang, Li, and Xiaojuan Huang. "Multiple SVM-RFE for multi-class gene selection on DNA microarray data." In 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1-6. IEEE, 2015. doi: https://doi.org/ 10.1109/IJCNN.2015.7280417.

[12] Naik, Akshata K., and Venkatanareshbabu Kuppili. "An embedded feature selection method based on generalized classifier neural network for cancer classification." Computers in Biology and Medicine 168 (2024): 107677. doi: https://doi.org/: 10.1016/j.compbiomed.2023.107677

[13] Mahendran, Nivedhitha, and Durai Raj Vincent PM. "A deep learning framework with an embedded-based feature selection approach for the early detection of the Alzheimer's disease." Computers in Biology and Medicine 141 (2022): 105056. doi: https://doi.org/10.1016/j.compbiomed.2021.105056

[14] Bommert, Andrea, Thomas Welchowski, Matthias Schmid, and Jörg Rahnenführer. "Benchmark of filter methods for feature selection in high-dimensional gene expression survival data." Briefings in Bioinformatics 23, no. 1 (2022): bbab354. doi: https://doi.org/10.1093/bib/bbab354

[15] Kundu, Rohit, Soham Chattopadhyay, Erik Cuevas, and Ram Sarkar. "AltWOA: Altruistic Whale Optimization Algorithm for feature selection on microarray datasets." Computers in biology and medicine 144 (2022): 105349. doi: https://doi.org/10.1016/j.compbiomed.2022.105349

[16] Guo, Yinan, Zirui Zhang, and Fengzhen Tang. "Feature selection with kernelized multi-class support vector machine." Pattern Recognition 117 (2021): 107988. doi: https://doi.org/10.1016/j.patcog.2021.107988

[17] Jain, V.K. Jain, R. Jain, An improved binary particle swarm optimization (ibpso) for gene selection and cancer classification using dna microarrays, in: 2018 Conference on Information and Communication Technology (CICT), IEEE, 2018, pp. 1–6. doi:https://doi.org/10.1109/INFOCOMTECH.2018.8722351

[18] Moradi, M. Gholampour, A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy, Appl. Soft Comput. 43 (2016) 117–130. doi: https://doi.org/10.1145/3321707.3321816

[19] Garibay, G. Sanchez-Ante, L.E. Falcon-Morales, H. Sossa, Modified binary inertial particle swarm optimization for gene selection in dna microarray data, in: Mexican Conference on Pattern Recognition, Springer, 2015, pp. 271–281. doi: https://doi.org/ 10.1007/978-3-319-19264-2_26

[20] Mohapatra and S. Chakravarty, "Modified PSO based feature selection for Microarray data classification," 2015 IEEE Power, Communication and Information Technology Conference (PCITC), Bhubaneswar, India, 2015, pp. 703-709. doi: https://doi.org/10.1109/PCITC.2015.7438088

[21] K.-H. Chen, K.-J. Wang, K.-M. Wang, M.-A. Angelia, Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data, Appl. Soft Comput. 24 (2014) 773–780. doi: https://doi.org/10.1016/j.asoc.2014.08.032

[22] Almugren, H. Alshamlan, Ff-svm: new firefly-based gene selection algorithm for microarray cancer classification, in: 2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE, 2019, pp. 1–6. doi: https://doi.org/10.1109/CIBCB.2019.8791236

[23] Jinthanasatian, S. Auephanwiriyakul, N. Theera-Umpon, Microarray data classification using neuro-fuzzy classifier with firefly algorithm, in: 2017 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2017, pp. 1–6. doi: https://doi.org/10.1109/SSCI.2017.8280967

[24] Raguntharr, S. Selvakumar, A wrapper based feature selection in bone marrow plasma cell gene expression data, Cluster Comput. 22 (6) (2019) 13785–13796. doi: https://doi.org/10.1007/s10586-018-2094-2

[25] M.S. Pratiwi, A. Aditsania., Cancer detection based on microarray data classification using genetic bee colony (gbc) and conjugate gradient backpropagation with modified polak ribiere (mbp-cgp), in: 2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA), IEEE, 2018, pp. 163–168 doi: https://doi.org/10.1109/IC3INA.2018.8629538

[26] M.A. Tawhid, A.M. Ibrahim, Feature selection based on rough set approach, wrapper approach, and binary whale optimization algorithm, Int. J. Machine Learn. Cybernet. 11 (3) (2020) 573–602. doi: https://doi.org/10.1007/s13042-019-00996-5

[27] Zakeri, A. Hokmabadi, Efficient feature selection method using real-valued grasshopper optimization algorithm, Expert Syst. Appl. 119 (2019) 61–72.doi: https://doi.org/10.1016/j.eswa.2018.10.021

[28] Chatra, V. Kuppili, D.R. Edla, A.K. Verma, Cancer data classification using binary bat optimization and extreme learning machine with a novel fitness function, Med. Biol. Eng. Comput. 57 (12) (2019) 2673–2682. doi: https://doi.org/10.1007/s11517-019-02043-5

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

[30] Allam, M. Nandhini, Optimal Feature Selection Using Binary Teaching Learning Based Optimization Algorithm, Journal of King Saud UniversityComputer and Information Sciences, 2018. doi: https://doi.org/10.1016/j.jksuci.2018.12.001

[31] Sharma, K.K. Paliwal, S. Imoto, S. Miyano, A feature selection method using improved regularized linear discriminant analysis, Mach. Vis. Appl. 25 (3) (2014) 775–786. doi: https://doi.org/10.1007/s00138-013-0577-y

[32] Roozbahani, Zahra, M. Yari, and Razieh Ghiasi. "Developing a Filter-Wrapper Feature Selection Method and its Application in Dimension Reduction of Gen Expression." Soft Computing Journal 6, no. 2 (2021): 48-59.

[33] Ke, Lin, Min Li, Lei Wang, Shaobo Deng, Jun Ye, and Xiang Yu. "Improved swarm-optimization-based filter-wrapper gene selection from microarray data for gene expression tumor classification." Pattern Analysis and Applications 26, no. 2 (2023): 455-472. doi: https://doi.org/10.1007/s10044-022-01117-9

[34]Jain, Rahi, and Wei Xu. "Artificial Intelligence based wrapper for high dimensional feature selection." BMC bioinformatics 24, no. 1 (2023): 392. doi: https://oi.org/10.1007/s00138-013-0577-y

[35] Song, Yu-Wei, Jie-Sheng Wang, Yu-Liang Qi, Yu-Cai Wang, Hao-Ming Song, and Yi-Peng Shang-Guan. "Serial filter-wrapper feature selection method with elite guided mutation strategy on cancer gene expression data." Artificial Intelligence Review 58, no. 4 (2025): 1-49. doi: https://doi.org/10.1007/s10462-024-11029-1

[36] Li, Zong-Zheng, Fang-Ling Wang, Feng Qin, Yusliza Binti Yusoff, and Azlan Mohd Zain. "Feature selection of gene expression data using a modified artificial fish swarm algorithm with population variation." IEEE Access 12 (2024): 72688-72706. doi:https://doi.org/10.1109/ACCESS.2024.3402652

[37]F.V. Sharbaf, S. Mosafer, M.H. Moattar, A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization, Genomics 107 (6) (2016) 231–238. doi: https://doi.org/10.1016/j.ygeno.2016.05.001

[38]Jain, V.K. Jain, R. Jain, Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification, Appl. Soft Comput. 62 (2018) 203–215. doi: https://doi.org/10.1016/j.asoc.2017.09.038

[39]Chinnaswamy, R. Srinivasan, Hybrid feature selection using correlation coefficient and particle swarm optimization on microarray gene expression data, in: Innovations in Bio-Inspired Computing and Applications, Springer, 2016, pp. 229–239. doi: https://doi.org/10.1007/978-3-319-28031-8_20

[40]Pashaei, M. Ozen, N. Aydin, Gene selection and classification approach for microarray data based on random forest ranking and bbha, in: 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), IEEE, 2016, pp. 308–311. doi: https://doi.org/10.1109/BHI.2016.7455896

[41]Dabba, A. Tari, S. Meftali, Hybridization of moth flame optimization algorithm and quantum computing for gene selection in microarray data, J. Ambi. Intell. Humanized Comput. (2020) 1–20. doi: https://doi.org/10.1007/s12652-020-02434-9

[42]S.K. Baliarsingh, S. Vipsita, B. Dash, A new optimal gene selection approach for cancer classification using enhanced jaya-based forest optimization algorithm, Neural Comput. Appl. 32 (12) (2020) 8599–8616. doi: https://doi.org/10.1007/s00521-019-04355-x

[43] Alanni, J. Hou, H. Azzawi, Y. Xiang, New gene selection method using gene expression programing approach on microarray data sets, in: International Conference on Computer and Information Science, Springer, 2018, pp. 17–31. doi: https://doi.org/10.1007/978-3-319-98693-7_2

[44] Xu, Zhaozhao, Fangyuan Yang, Chaosheng Tang, Hong Wang, Shuihua Wang, Junding Sun, and Yudong Zhang. "FG-HFS: A feature filter and group evolution hybrid feature selection algorithm for high-dimensional gene expression data." Expert Systems with Applications 245 (2024): 123069. doi: https://doi.org/10.1016/j.eswa.2023.123069

[45] Song, Yu-Wei, Jie-Sheng Wang, Yu-Liang Qi, Yu-Cai Wang, Hao-Ming Song, and Yi-Peng Shang-Guan. "Serial filter-wrapper feature selection method with elite guided mutation strategy on cancer gene expression data." Artificial Intelligence Review 58, no. 4 (2025): 1-49.doi: https://doi.org/ 10.1007/s10462-024-11029-1

[46] Yaqoob, Abrar, Navneet Kumar Verma, Rabia Musheer Aziz, and Mohd Asif Shah. "Optimizing cancer classification: a hybrid RDO-XGBoost approach for feature selection and predictive insights." Cancer Immunology, Immunotherapy 73, no. 12 (2024): 261. doi: https://doi.org/10.1007/s00262-024-03843-x

[47] Gong, Huanhuan, Yanying Li, Jiaoni Zhang, Baoshuang Zhang, and Xialin Wang. "A new filter feature selection algorithm for classification task by ensembling pearson correlation coefficient and mutual information." Engineering Applications of Artificial Intelligence 131 (2024): 107865. doi: https://doi.org/10.1145/3712199

[48]Gulande, Punam, and Raval Awale. "A Hybrid mRMR-RSA Feature Selection Approach for Lung Cancer Diagnosis Using Gene Expression Data." Biomedical and Pharmacology Journal 18, no. March Spl Edition (2025): 257-270. doi: https://doi.org/10.13005/bpj/3086

[49]Yaqoob, Abrar. "Combining the mRMR technique with the Northern Goshawk Algorithm (NGHA) to choose genes for cancer classification." International Journal of Information Technology (2024): 1-12. doi: https://doi.org/10.1007/s41870-024-01849-3

[50]Khaing, May Myat Myat, May Mar Oo, and Htoo Naing Aung. "Cancer Type Detection based on Gene Expression Data using Support Vector Machine." In 2024 Conference of Young Researchers in Electrical and Electronic Engineering (ElCon), pp. 230-234. IEEE, 2024. doi: https://doi.org/10.1109/ElCon61730.2024.10468149

[51]Peng, Hanchuan, Fuhui Long, and Chris Ding. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy." IEEE Transactions on pattern analysis and machine intelligence 27, no. 8 (2005): 1226-1238. doi: https://doi.org/10.1109/TPAMI.2005.159

[52]Ding, Chris, and Hanchuan Peng. "Minimum redundancy feature selection from microarray gene expression data." Journal of bioinformatics and computational biology 3, no. 02 (2005): 185-205. doi: https://doi.org/10.1142/S0219720005001004

[53]Ramírez-Gallego, Sergio, Iago Lastra, David Martínez-Rego, Verónica Bolón-Canedo, José Manuel Benítez, Francisco Herrera, and Amparo Alonso-Betanzos. "Fast-mRMR: Fast minimum redundancy maximum relevance algorithm for high-dimensional big data." International Journal of Intelligent Systems 32, no. 2 (2017): 134-152. doi:10.1002/int.21833

[54] Mohammed, Mohanad, Henry Mwambi, Innocent B. Mboya, Murtada K. Elbashir, and Bernard Omolo. "A stacking ensemble deep learning approach to cancer type classification based on TCGA data." Scientific reports 11, no. 1 (2021): 15626. Doi: 10.1038/s41598-021-95128-x

[55] Divate, Mayur, Aayush Tyagi, Derek J. Richard, Prathosh A. Prasad, Harsha Gowda, and Shivashankar H. Nagaraj. "Deep learning-based pan-cancer classification model reveals tissue-of-origin specific gene expression signatures." Cancers 14, no. 5 (2022): 1185. doi: 10.3390/cancers14051185

[56] Ramirez, Ricardo, Yu-Chiao Chiu, Allen Hererra, Milad Mostavi, Joshua Ramirez, Yidong Chen, Yufei Huang, and Yu-Fang Jin. "Classification of cancer types using graph convolutional neural networks." Frontiers in physics 8 (2020): 203. doi: 10.3389/fphy.2020.00203