Int J Cancer Manag. 2023 December; 16(1):e135724. https://doi.org/10.5812/ijcm-135724.

Published online 2023 May 17. **Research Article**

Classification of Potential Breast/Colorectal Cancer Cases Using Machine Learning Methods

Maryam Jafarpour [1], Ali Moeini [1,*], Niloofar Maryami [2], Azin Nahvijou [3] and Ayoub Mohammadian [4]

[1] Department of Algorithms and Computation, School of Engineering Sciences, College of Engineering, University of Tehran, Tehran, Iran
[2] Student Research Committee, Zanjan University of Medical Sciences, Zanjan, Iran
[3] Cancer Research Center, Cancer Institute of Iran, Tehran University of Medical Sciences, Tehran, Iran
[4] Department of Information Technology Management, Faculty of Management, University of Tehran, Iran

[*] *Corresponding author*: Department of Algorithms and Computation, School of Engineering Sciences, College of Engineering, University of Tehran, Tehran, Iran. Email: moeini@ut.ac.ir

Received 2023 March 07; **Accepted** 2023 March 18.

Abstract

Background: The algorithmic classification of infected and healthy individuals by gene expression has been a topic of interest to researchers in numerous domains, including cancer. Several studies have presented numerous solutions, such as neural networks and support vector machines (SVMs), to classify a diverse range of cancer cases. Such classifications have provided some degrees of accuracy, which highly depend on optimization approaches and suitable kernels.

Objectives: This study aimed at proposing a method to classify cancer-prone and healthy cases under breast cancer and colorectal cancer (CRC), using machine learning methods efficiently, increasing the accuracy of the classification process.

Methods: This study presented an algorithm to diagnose individuals prone to breast cancer and CRC. The novelty of this algorithm lies in its suitable kernel and the feature extraction approach. By the application of this algorithm, this study first identified the genes closely associated with these types of cancers and, then, tried to find individuals susceptible to the concerned cancers using SVM. The present study highlighted the indirect gene expressions associated with these cancers, which might show health status complications for the patients. To this end, the algorithm consists of SVMs in conjunction with the k-fold method for validation.

Results: The results confirmed the superior performance of this approach, compared to the common neural networks. The algorithm's identification accuracy values were 98.077% and 99.806% for breast cancer and CRC, respectively. The graphic representation of the cause-effect relationships was also provided to help researchers better understand the trend of cancer or other types of diseases.

Conclusions: The feature extraction method highly affects the accuracy of the classification. In addition, relying on indirect disease-triggering genes' expressions highlights a cause-effect relationship between genes and diseases. Such relationships can form Markov models in the clinical domain leading to treatment paths and prediction of patient outcomes.

Keywords: Machine Learning, Support Vector Machine, Gene Expression, Breast Cancer, Colorectal Cancer, Classification

1. Background

Abundant genetic data on the use of conventional methods have failed to find the relationship between genes and cancer (or any other type of disease). Artificial intelligence and machine learning are critical tools in discovering such relationships (1). Diagnosing individuals prone to cancer is one of the machine learning applications that this study pursued.

Numerous studies have focused on the diagnosis and treatment of various cancers, including oral, pediatric, melanoma, lung, and gastric cancers, using patients' genetic data (2-4). Breast cancer and colorectal cancer (CRC) have also been a hotbed of research. In this regard, about 19.3 million new cancer cases (18.1 million excluding non-melanoma skin cancer) and almost 10.0 million cancer deaths (9.9 million excluding non-melanoma skin cancer) were reported worldwide in 2020. Female breast cancer surpassed lung cancer as the most commonly diagnosed cancer, with about 2.3 million new cases (11.7%), followed by lung (11.4%), colorectal (10.0%), prostate (7.3%), and stomach (5.6%) cancers. Lung cancer remained the leading cause of cancer death, accounting for about 1.8 million deaths (18%), followed by colorectal (9.4%), liver (8.3%), stomach (7.7%), and female breast (6.9%) cancers

(5). In addition to the prevalence of cancer, any patient diagnosed with cancer burdens a high cost to the family and society to get proper treatment. Such a treatment varies largely by tumor type and stage (6) or even the care provider facilities regarding the number of costs to be covered (7). According to the Financial Burden of Cancer calculated for the US, on average for all cancer sites, this country approximately spent $208.9 billion to treat patients in 2020. This is $29.8 billion for female breast cancer with the highest costs followed by colorectal cancer with $24.3 billion (8). Such an economic burden forces a financial toxicity that may lead to lower quality of life, general loss of productivity, financial debt, etc. Cancers may rise due to many causes including tobacco, pharmaceuticals, hormones, human immunodeficiency virus, etc. (9). At the genetic level, these factors may lead to a change in the expression level of some genes.

So far, some genes, including *MDM2, CD82, MED1, miR-34a, miR-520h, HDAC1, CYP1A1, NTN4, EIF4EBP1, ATG4A, BAG1, MAP1LC3A, SERPINA1, MMP1*, and *MMP9* are identified as the risk factors of breast cancer. Relevant studies on these genes concluded that they could develop breast cancer (10-17). Similarly, some genes, including *PTEN, P62, NOD2, TP53, MSH3, POFUT1, RPRD1B, EIF6, MGP, FGF2, OGDHL, CYP24A1, WNT1, KLF5, WNT16, CLDN1*, and *TET3* have been effective in CRC development (18-25).

Analyzing any change in the expression of the genes can help in understanding the main causes of the disease and reveals at which state the disease is and how to handle it and plan proper and efficient treatments in numerous fields of application, such as drug designs (26) and personalized medicines (27).

The gene expressions have been studied, using different methods, including heat maps, clustering, gene set enrichment, and pathway analysis using gene ontology, network analysis, and machine learning (28-30). Machine learning is used to predict information, including survival rates, outcomes, and disease diagnosis (30) that lies in supervised and unsupervised categories. As a supervised learning approach, SVM with a wide variety of kernels has been adopted in the biomedical domain including breast cancer, CRC, etc. (31-34).

The studies were conducted from different viewpoints with different datasets from structured to unstructured data. For instance, de Ronde et al. (33) used SVM to predict breast cancer chemotherapy resistance. Although their results did not show a huge variance of accuracy on the algorithm; however, the results for SVM might improve by adopting some other kernels or feature selection methods. In another study, Smolander et al. (34) compared deep belief networks to SVMs to classify gene expression data in deep learning and traditional machine learning approaches. They realized that combining deep belief networks and SVM outperforms traditional approaches.

In general, SVM demonstrates some advantages in many ways, specifically in high classification accuracy and small computation (35). In addition, one critical issue while using SVM is feature selection. Selecting the optimum number of features has a critical role especially when working with large datasets. This can speed up training, avoid overfitting and result in better classification. An improper set of features leads to inaccuracy of the classification result (35).

An issue with microarray data is that they are unbalanced; the number of available samples in each class is not equal, which makes the classification biased toward the class having the majority of samples; also, ranking of features is considered a challenge (36).

Feature selection is primarily focused on removing non-informative or redundant predictors from the model (37) that are either a wrapper or filter methods. Filter-type methods select variables regardless of the model relying on statistical methods. Wrapper methods evaluate subsets of variables, which allows for the detection of the possible interactions amongst variables (38). Thus, adopting a proper filter selection method can boost the output and performance of the classifier.

This study presents an SVM-based algorithm to diagnose individuals prone to breast cancer and CRC. This study has chosen an SMO solver as the kernel and followed a filter type of feature extraction. To achieve better and more accurate results, the k-fold method was used, which is a validation technique splitting the data into k subsets. Then, the average error from all these k trials forms the overall error. This is more reliable than the standard handout method.

2. Objectives

This study aimed at proposing a method to classify cancer-prone and healthy individuals under breast cancer and CRC, using machine learning methods efficiently, increasing the accuracy of the classification process. Such a classifier can, then, be used as an assistive tool for healthcare providers specifically pathologists and genetics scientists to diagnose patients with more accuracy and move toward better therapeutic solutions based on the set of selected features.

3. Methods

3.1. Data Sources

Different studies have used different datasets to evaluate their algorithms and obtain results. This study

used gene expression data. Gene expression datasets obtained from the National Center for Biotechnology Information (NCBI) database for breast cancer and CRC were GSE15852 and CRC GSE44076, respectively (39). The breast cancer dataset included gene expressions from 43 healthy and 43 tumor samples. The CRC dataset also included gene expressions from 148 healthy and 98 tumor samples. There were 22 283 and 49 386 genes in the datasets of breast cancer and CRC, respectively. The datasets were downloaded on March 2021. Their implementation and data analysis was performed, using Matlab 2017. The source code of the implemented algorithm is available on GitHub.

3.2. Feature Extraction

The large volume of information and the large dimensions of the problem make the machine learning algorithms first look for reducing problem dimensions. In machine learning algorithms, this section, known as feature selection or feature extraction, is one of the main steps in classification algorithms, upon which the computational/memory complexity and accuracy of the algorithms depend. The selection of numerous features increases the complexity of the algorithm and the accuracy of the algorithm and vice versa. Accordingly, the number of selected features is a trade-off between the computational/memory complexity and the algorithm's accuracy. In this study, the selected features were genes. The primary objective of this study was to select genes, whose changes in gene expression indicate the presence of cancer.

First, there should be a calculation of the average gene expression for all healthy and tumor samples to select the best features. It is necessary to remove the outlier data for the average calculation; however, due to the low variance, there was no need to remove any data. The formula for calculating these average numbers for each gene g are shown in Equations 1 - 4 where BCH_g (i) is the gene expression for the i'th healthy sample, $\widehat{BCH_g}$ is the mean value of the gene g expression for the healthy samples of breast cancer, and $\widehat{BCT_g}$ represents the mean value of the gene g expression for the tumor samples of breast cancer. The same is true for CRC.

$$\widehat{BCH_g} = \frac{\sum_{i=1}^{n} BCH_g(i)}{n} \quad (1)$$

$$\widehat{BCT_g} = \frac{\sum_{i=1}^{n} BCT_g(i)}{n} \quad (2)$$

$$\widehat{CRCH_g} = \frac{\sum_{i=1}^{n} CRCH_g(i)}{n} \quad (3)$$

$$\widehat{CRCT_g} = \frac{\sum_{i=1}^{n} CRCT_g(i)}{n} \quad (4)$$

After calculating the mean values, the best features were selected from the genes for data clustering. To this end, the average ratio of each gene in the healthy state to the tumor state was calculated according to Equations 5 and 6.

$$BC_g = \frac{\widehat{BCH_g}}{\widehat{BCT_g}} \quad (5)$$

$$CRC_g = \frac{\widehat{CRCH_g}}{\widehat{CRCT_g}} \quad (6)$$

Now each gene should have a score to select the best genes based on this score. The scores were assigned in a way that any decrease or increase in gene expression would have the same effect. With this limitation, Equations 7 and 8 are the best to score the genes. The absolute value of the logarithm causes the ratios $\frac{a}{b}$ and $\frac{b}{a}$ to get the same scores, which is desirable for this study.

$$BC - Score_g = |logBC_g| \quad (7)$$

$$CRC - Score_g = |logCRC_g| \quad (8)$$

3.3. Support Vector Machine

As previously mentioned, this study clustered the samples into two categories of normal and cancer-prone. Accordingly, SVM was useful in this classification. This study used the SVM with the SMO solver approach for the quadratic kernel (40). The SVM relies on a classifier depending on the dataset to be linear for two-dimensional ones or any hyperplane for multidimensional. The SMO solver boosts the training process more efficiently.

The present study also used a neural network to investigate the performance of SVM. Figure 1 depicts the general structure of the machine learning method and the prediction based on supervised learning. The output model after training is a trained SVM or a trained neural network.

3.4. Validation

To examine the performance of a trained machine, a part of the data is to train the machine, and another part is to test its performance or so-called accuracy calculation. Nevertheless, the part of the data that is for training is highly important because the machine might be well-trained with some parts of the data and not by some other parts. To solve this problem, this study used the k-fold method, in which the data are divided into k sections. At each step, k-1 parts of the data

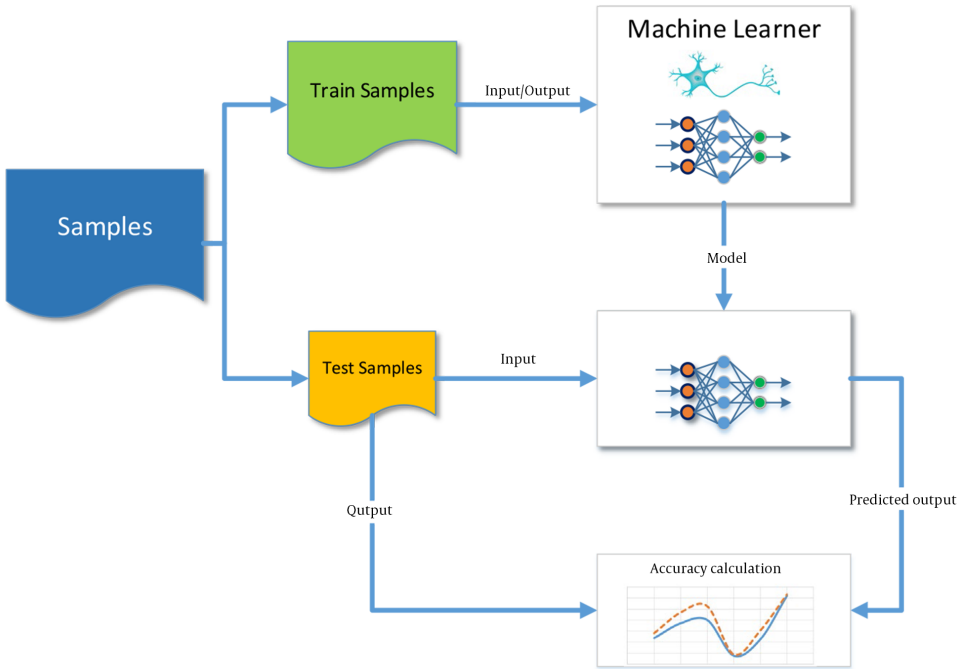

Figure 1. Machine-learning flowchart for supervised learning

are for training, and one part is to test the machine's performance. Accordingly, the number of steps to check the accuracy of the performance is equal to k. Figure 2 shows the sections used for training and testing with k = 4. This method's final accuracy equals the average accuracy calculated at all steps.

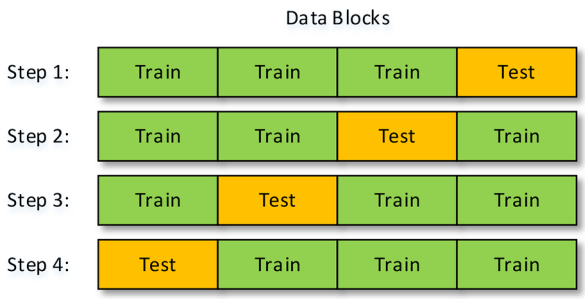

Figure 2. K-fold steps with k = 4

4. Results

In this study, 20 features were extracted from each cancer dataset, using the feature extraction method. Tables 1 and 2 show the results of the absolute logarithmic values of the extracted genes for CRC and breast cancer, respectively. Tables 3 and 4 show a list of the genes introduced in previous CRC and breast cancer studies, respectively.

As illustrated in Figures 3 to 6 and Tables 1 and 2, the gene expression levels for the genes obtained in this study show a significant difference, compared to the genes introduced in previous studies (as an indicator of breast cancer or colorectal cancer). This finding can indicate that the genes in this study are linked to colorectal and breast cancers more closely or they might be an underlying cause or a complication of the disease. This is discussed in more detail in the Discussions section. Table 5 shows that this algorithm detected CRC by more than 99%, using both neural networks and SVM. Nevertheless, the aforementioned results are less accurate for breast cancer. This might be due to several reasons, including more diversity in the gene expressions of breast cancer. However, diagnosing breast cancer using SVM is still more acceptable than using neural networks. In this study, the false positives are of great importance because a method to identify the susceptible cases was sought as much as possible. Accordingly, reducing the false positives indicates a decrease in the error rate of the current proposed approach. Any algorithm and its optimizations might have an error rate. A combination of different

Table 1. Absolute Logarithmic Ratio in Colorectal Cancer Samples for Genes Selected in This Study

Gene Symbol	Normal Average	Tumor Average	Absolute Logarithmic Ratio
FOXQ1	2.223910908	6.897400898	0.491568084
SFRP1	7.791344622	2.937684327	0.423607288
PLP1	6.502352959	2.52659399	0.410535081
ABCG2	8.859031714	3.351576673	0.422137097
COL11A1	2.197618612	6.152669908	0.447111291
AQP8	10.45230679	3.832825541	0.435693096
CEACAM7	8.566520765	3.396245367	0.401805413
CA1	10.78996539	3.816095082	0.451400865
CA1	12.30637631	4.583258724	0.428955817
CLDN8	6.599359908	2.4096035	0.437556229
CD177	8.290931235	3.018891276	0.438755841
GUCA2B	10.2922992	3.742539765	0.439345979
CA7	7.444865847	2.79510298	0.425459063
BMP3	5.524542541	2.158379102	0.408168595
TMIGD1	10.06301872	3.132877398	0.506784881
SLC4A4	7.336944143	2.864589408	0.408452831
MMP7	2.076586173	7.631830949	0.565278784
KRT23	2.619680918	7.065921806	0.43092043
ADH1B	8.658761071	3.07382252	0.449776968
OTOP2	7.890946133	2.733458888	0.460416532

Table 2. Absolute Logarithmic Ratio in Breast Cancer Samples for Genes Selected in This Study

Gene Symbol	Normal Average	Tumor Average	Absolute Logarithmic Ratio
CDH1	10.12790132	11.38846291	0.05095
KRT19	8.771923697	10.43900939	0.07556
LPL	12.19861268	10.57707802	0.06194
CFD	12.98469738	11.51893758	0.05202
ADIPOQ	12.56529143	10.92177526	0.06088
HBB	12.1307117	10.26378344	0.07258
AKR1C3	10.54135091	9.341985121	0.05246
CD36	12.3276894	10.8627203	0.05494
ADH1B	12.4294803	10.89986437	0.05703
GYG2	11.50944658	9.867195347	0.06686
SORBS1	10.83684142	9.318539532	0.06555
RBP4	11.31251754	9.626174305	0.07011
PCOLCE2	10.3687605	9.14412402	0.05458
CD24	9.35284475	11.19384905	0.07804
ACACB	11.51449992	10.21201579	0.05213
CDH1	10.12790132	11.38846291	0.05095

Table 3. Absolute Logarithmic Ratio in Colorectal Cancer Samples for Genes Introduced in Previous Studies

Gene Symbol	Normal Average	Tumor Average	Absolute Logarithmic Ratio
RPRD1B	6.340062038	7.162458816	0.052968630
OGDHL	3.160723867	3.047873622	0.015789601
FGF2	3.554201888	2.781068847	0.106530353
TET3	6.395727755	6.383841133	0.000807898
MSH3	3.803539776	4.019127286	0.023943798
WNT16	1.999506153	1.973441102	0.005698576
CYP24A1	2.918686673	2.944265898	0.003789554
WNT1	2.182406765	2.191098184	0.001726140
CLDN16	2.279791786	2.307774643	0.005298213
POFUT1	3.736163612	4.471701408	0.078046910
KLF5	10.77017785	10.42091324	0.014317094
NOD2	2.392484296	2.436204337	0.007864616
EIF6	6.735043245	7.590437622	0.051926427
PTEN	7.299340194	6.935050888	0.022233953
MGP	4.293408755	3.166209949	0.132262528
TP53	2.683255867	2.683807286	0.000089239

Table 4. Absolute Logarithmic Ratio in Breast Cancer Samples for Genes Introduced in Previous Studies

Gene Symbol	Normal Average	Tumor Average	Absolute Logarithmic Ratio
HDAC1	10.56569313	10.79593658	0.00936
BAG1	11.2298917	11.18956404	0.00156
MED1	9.485055809	9.657937687	0.00784
CD82	11.2675989	11.28307914	0.00060
MMP9	10.5577457	10.66589147	0.00443
MMP1	8.939618655	9.223201462	0.01356
MDM2	9.195558765	9.183370978	0.00058
CYP1A1	11.07315899	11.02370657	0.00194
SERPINA1	9.68937968	9.706017316	0.00075
ATG4A	10.94181449	10.96255842	0.00082
EIF4EBP1	11.64878392	11.61895678	0.00111
HDAC1	10.56569313	10.79593658	0.00936
BAG1	11.2298917	11.18956404	0.00156
MED1	9.485055809	9.657937687	0.00784
CD82	11.2675989	11.28307914	0.00060
MMP9	10.5577457	10.66589147	0.00443

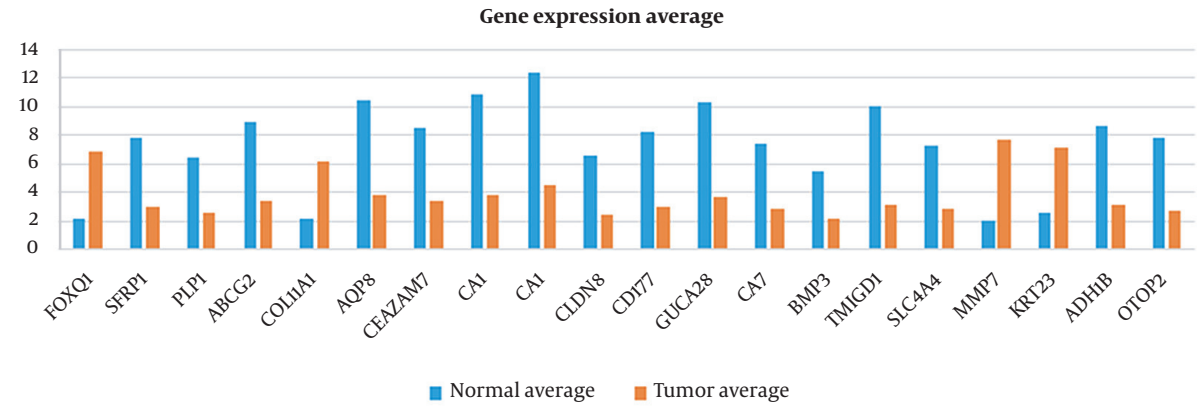

Figure 3. Gene expression average in normal and tumor colorectal cancer samples for genes selected in this study

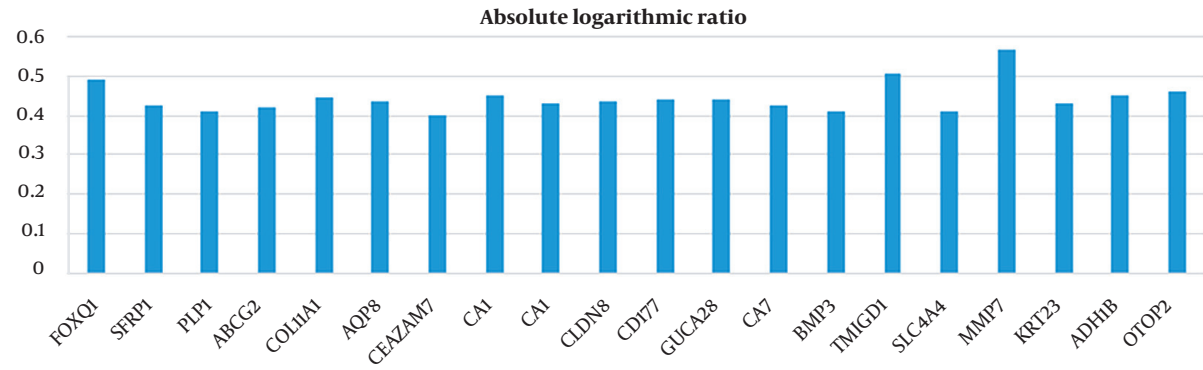

Figure 4. Absolute logarithmic ratio of colorectal cancer samples for genes selected in this study

algorithms with different error rates would lead to the cross-product of their error rates, which would logically lead to a lower error rate. Using the k-fold technique helped reduce this rate to the lowest rate possible in the current proposed approach.

Table 5. Algorithm Accuracy in Classification of Individuals with Tumors and Healthy Individuals Using Neural Network and Support Vector Machine

	Neural Network, %	Support Vector Machine, %
Breast cancer	85.385	98.077
Colorectal cancer	99.675	99.806

5. Discussion

This study presented a novel approach to classify breast cancer and CRC cases. Moreover, to ensure that the training process is proper and efficient, the k-fold method was followed. This study used the gene expression datasets for breast cancer and CRC (namely breast cancer GSE15852 and CRC GSE44076) obtained from the NCBI database. The datasets included gene expressions from 43 healthy and 43 tumor samples for breast cancer and 148 healthy and 98 tumor samples for CRC, respectively. The findings indicated that the proposed SVM-based approach, in conjunction with the k-fold method, performs much better than neural networks, with accuracy values of 98.077% and 99.806% for breast cancer and CRC, respectively. This result is highly impressive, compared to the results of similar studies, and is due to the way the extraction of the features (33, 34, 41, 42).

For instance, de Ronde et al.'s (33) study aimed at evaluating the performance of subtype-specific and non-specific predictors. They studied several approaches and compared their performance including SVM. A deeper look into the findings showed that for subtype predictors based on gene expression data, the accuracy

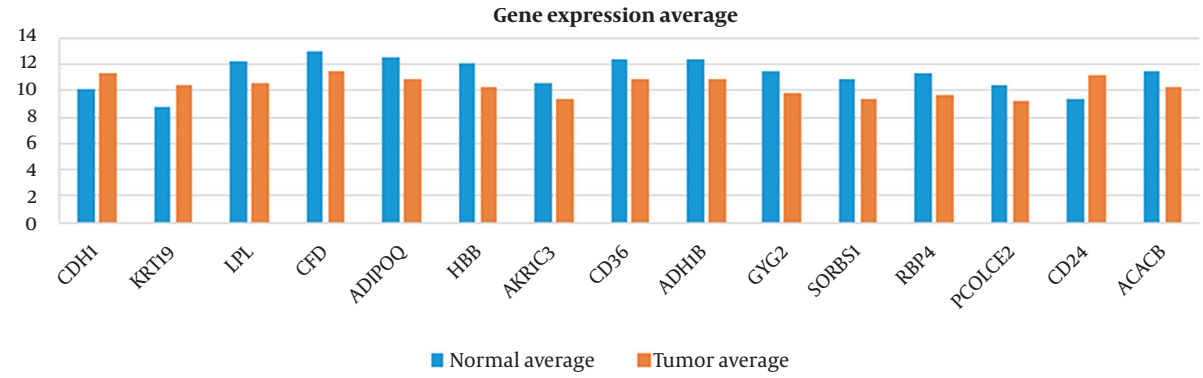

Figure 5. Gene expression average in normal and tumor breast cancer samples for genes selected in this study

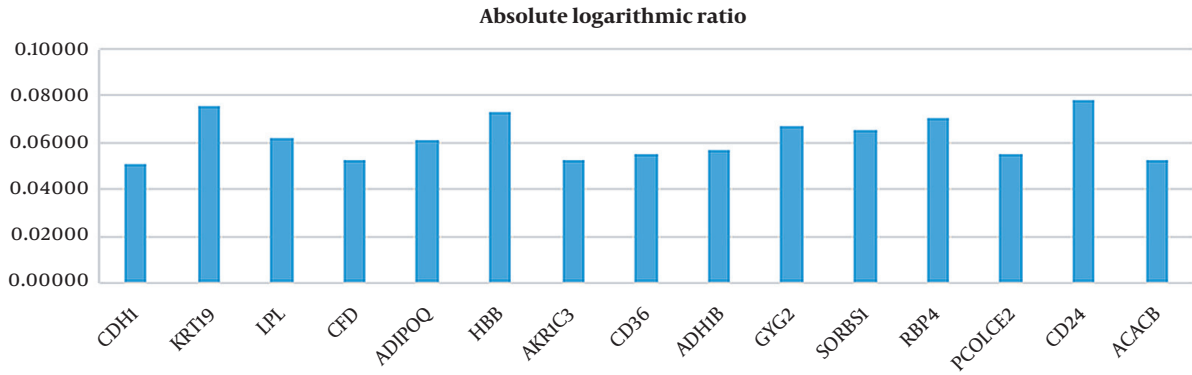

Figure 6. Absolute logarithmic ratio of breast cancer samples for genes selected in this study

values of the results were 76.22% and 75.59% for the SVM-Wilcoxon-Mann-Whitney (SVM+WMW) and the SVM-Wilcoxon-Mann-Whitney with uncorrelated features (SVM+WMW uncor.), respectively. This finding shows that our proposed approach performed much better regarding the classification of breast cancer. Furthermore, the most outstanding result reported by Smolander et al. (34) was the utilization of the backpropagation algorithm in conjunction with SVM, resulting in an accuracy of 90.16% for breast cancer in comparison to which the present study revealed a much better performance and accuracy of 98.077%. The reason for such a difference in the accuracy of their study with the present paper is that they selected the features with the highest variance to perform the classification task. Relying on the features with high variance indicates that there is more variation on the data values that may increase the error rate. To overcome this issue, the authors have used back propagation algorithm to handle such variations and reduce the error rate. However, the accuracy remained lower that the one achieved by our study. The same is true for other similar studies by Chiu et al. (41) and Xu et al. (42).

Many other studies have adopted several approaches to classify breast cancer cases, using different methods from supervised binary classification to unsupervised deep learning methods.

Egwom et al. (43) used SVM along with linear discriminant analysis (LDA). LDA is a technique for dimensionality reduction. The concept of dimensionality reduction may intuitively have similarities with feature selection, but the two techniques are different.

Feature selection and dimensionality reduction are often grouped. While both methods are used for reducing the number of features in a dataset, there is an important difference.

Feature selection is simply selecting and excluding given features without changing them. Dimensionality reduction transforms features into a lower dimension.

The dataset used in this paper is different from what we did. This difference can be examined from several aspects. One is related to the type of data used. In the mentioned article, the Wisconsin Breast Cancer dataset (WBCD) contains information and characteristics of the tumor, which has 569 records. The next dataset is the Wisconsin Prognostic Breast Cancer dataset (WPBC), which has 198 records and contains breast cancer tumor characteristics and information. One of the main characteristics and the fundamental difference between our study and the mentioned article is that our dataset is fundamentally different in terms of content and we are dealing with gene expression. On the other hand, our data contains a wide range of genes identified in patient samples, and this causes the number of features we deal with to be much higher than the number of features in the above datasets. The accuracy of their proposed algorithm was 99.2% and 79.5% for WBCD and WPBC datasets, respectively. This variation of accuracy lies under two reasons first variation of features for these two datasets and second large variance of values for each dataset. Although their results cannot be compared to ours because of the reasons above, this result also emphasized that SVM is the proper tool for binary classification, and combining it with proper feature selection or dimensionality reduction can improve its performance (43).

In a similar study by Naji et al. (44), the dataset used in it is different from what we did. This difference can be examined from several aspects. One is related to the type of data used. In the mentioned article, WBCD is used (same as the study by Egwom et al. (43)).

However, the accuracy we obtained is very similar to the accuracy reported in this paper. On the other hand, similar to our findings, in this article, it has been concluded that the performance of SVM is better for binary classifications.

The paper by Aljuaid et al. has achieved acceptable results such that a very interesting accuracy of 99.7% has been achieved for the ResNet classifier for the classification of clinical images related to breast cancer using the deep learning method. This approach is different from the supervised learning approach we were looking for, and therefore no concrete comparison can be made in this regard (45).

The data source used in the study by Arooj et al. is ultrasound images and histopathology images, and the CNN-AlexNet model is also used, which is a combination of deep learning. The accuracy of this model has varied between 96.07 and 100% depending on the different datasets used (46).

The most similar research to our work so far is the work by Wu and Hicks The tool they used was R. In their study, SVM achieved the highest accuracy. But, since our model has been used for both breast cancer and CRC classification, the accuracy of our model has decreased for breast cancer classification (47). However, it is still very acceptable.

Therefore, utilizing a proper feature selection method presents much better results in SVM-based approaches for either machine learning or deep learning (34, 48). This study adopted an approach to be used for machine learning feature selection, in which the gene expression levels have changed the most. The average gene expressions have provided us with proper features and led us to identify some implicit information regarding the potential variations in gene expressions for the infected and healthy individuals.

In other words, these genes might not necessarily indicate the direct cause of infection by the disease; however, these genes show a potential to suspect the infected cases. These genes might not be the ones triggering the cancer incident; nevertheless, cancer-prone cases might affect these genes. They might also indicate adverse reactions to, for example, drug use and therapy.

It is noteworthy that the above-mentioned results for the set of selected features can be the cause/effect of the disease.

For instance, the current study has selected *SFRP1* (Table 1), a suppressor gene located in a chromosomal region frequently deleted in breast cancer (49). Another study revealed that *SFRP1* gene methylation in CRC was associated with lymph node invasion (50). On the other hand, *PTEN* has been widely studied as an indicator of CRC (18). A deeper look at the Tables 1 and 2 shows that relying on the set of genes that have already been identified as tumor markers (i.e. *PTEN* in Table 3) may mislead us based on our dataset. Because they might not show a significant difference between healthy and tumor cases, leading to inaccurate classification.

The same can be concluded for breast cancer. For instance, *MMP9* has been studied as the risk factor for breast cancer (17). However, according to Table 4, the expression of this gene in our dataset could not differentiate the healthy and tumor case because of the close normal average and tumor average. Instead, *ADIPOQ* has been identified as one of the proper features in our dataset. *ADIPOQ* does not trigger breast cancer but it is shown that it increases the efficacy of chemotherapeutic agents. Notably, high expression of *ADIPOQ* receptor *ADIPOR2*, *ADIPOQ*/adiponectin, and *BECN1* significantly correlate with increased overall survival in chemotherapy-treated breast cancer patients (51). This is shown in Figures 7 - 10 visually.

Accordingly, it can be concluded that the expression of

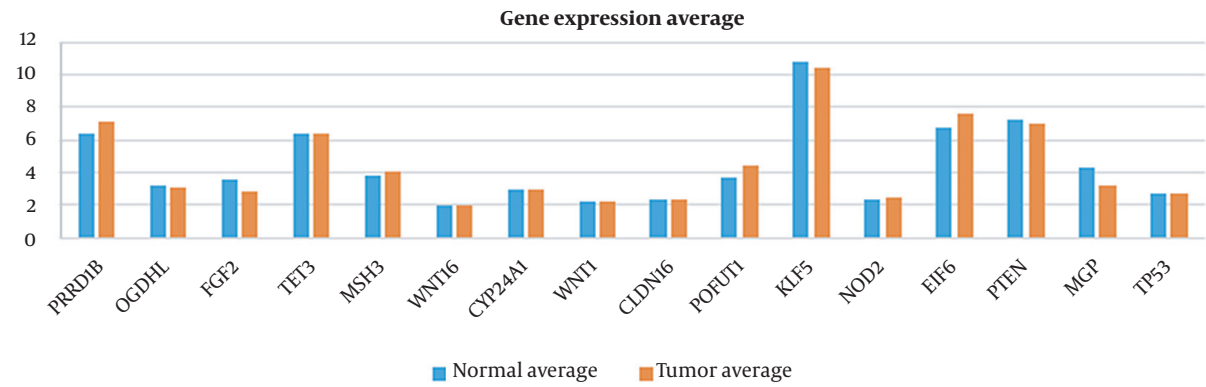

Figure 7. Gene expression average in normal and tumor colorectal cancer samples for genes introduced in previous studies

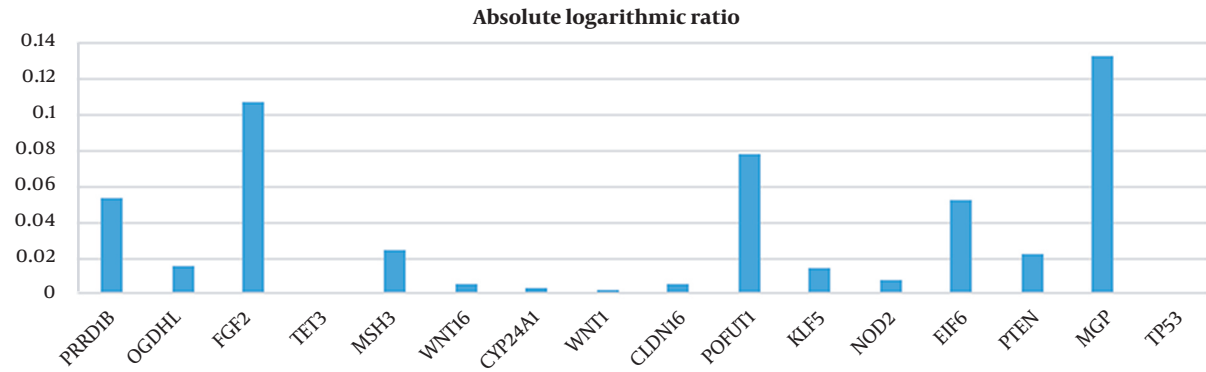

Figure 8. Absolute logarithmic ratio of colorectal cancer samples for genes introduced in previous studies

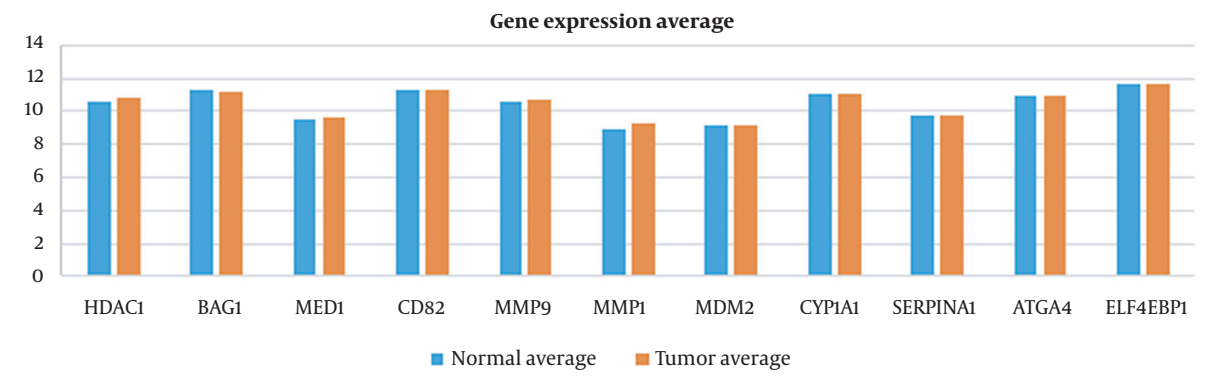

Figure 9. Gene expression average in normal and tumor breast cancer samples for genes introduced in previous studies

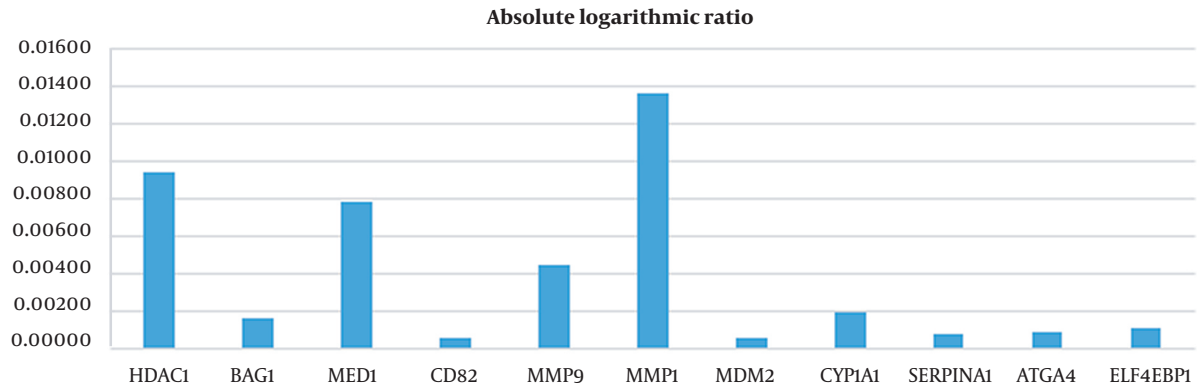

Figure 10. Absolute logarithmic ratio of breast cancer samples for genes introduced in previous studies

some genes, such as *MSH3* (for breast cancer) and *SFRP1* (for CRC) has effects on the disease progression; however, some others detected in the present study might not be known yet, or they might not be studied extensively to have any direct impact on the concerned diseases, yet. As previously mentioned, such effects can be studied further as causes and effects.

A directed acyclic graph (DAG) or directed cyclic graph (DCG) can visualize such effects. The DAGs are used to model a priori causal assumptions (52). In some cases, the relationships might be mutual (i.e., any change in the expression of a certain gene might cause the loss or overexpression of another gene). The cyclic relationship might occur either directly or indirectly (Figure 11). In future studies, visualizing such relationships among the genes in this study can reveal valuable implicit information. Furthermore, the graphs lead us to a better understanding of the chain of transformations or any other disease and might help manage the disease by cutting these transformation chains.

In Figure 11A, it is assumed that a change in the expression level or the mutation of gene A (in terms of loss or overexpression) affects the expression or mutation of genes B, C, and F. Gene C itself affects gene E, which affects genes F and G. In this case, all the relationships form a tree (or an acyclic graph). However, in Figure 11B, any change in the expression or mutation of gene G might affect gene A. Visualizing such a relationship would lead to a DCG.

The graphs in Figure 11 can also be considered weighted graphs. In this model, the weight of the edges of the graph indicates the probability of the effect of a change in the level of gene expression or the mutation of one gene on the gene expression or mutation of other genes. Figure 12 shows an example of this type of weighted graph. This graph is a Markov chain because it has the Markov property (i.e., being memoryless because the conditions at the time [step] t+1 only depend on the inputs and conditions at the time [step] t). Equation 9 shows the formal definition of this property, where X is a discrete random variable.

$$Pr\left(X_{t+1} = x | X_1 = x_1, X_2 = x_2, \ldots, X_t = x_t\right) \quad (9)$$

$$= Pr\left(X_{t+1} = x | X_t = x_t\right)$$

Considering the Markov chain as a causal model, the capabilities of the Markov chain (e.g., the steady state) can be used to investigate some issues in the medical field. One can consider pharmaceutical products and medical care (e.g., surgery) inputs in this causal system. Accordingly, using the steady state, it is possible to predict which direction the patient's condition would eventually go by the use of special medicines and care (i.e., recovery, further complications, or death).

Our proposed approach gave us a set of genes that are not necessarily the main causes of the disease; however, they may be the side effects of the disease. In other words, the changes in the expression of the genes responsible for breast cancer and CRC have led to a change in the expression level of a set of genes in different patients, which indicates the occurrence of complications caused by the disease or other changes and developments in such patients. Since these affected genes were more apparent in the dataset, the proposed model of this study obtained better results.

Another advantage of this approach is that it identifies a diverse range of potential gene expressions affected by the diseases, which is the superior capability of SVM. Not all the identified gene expressions in this paper (as features) trigger the disease itself; however, the study of their effects on the disease and the study of their relationships might further lead to new observations and even new

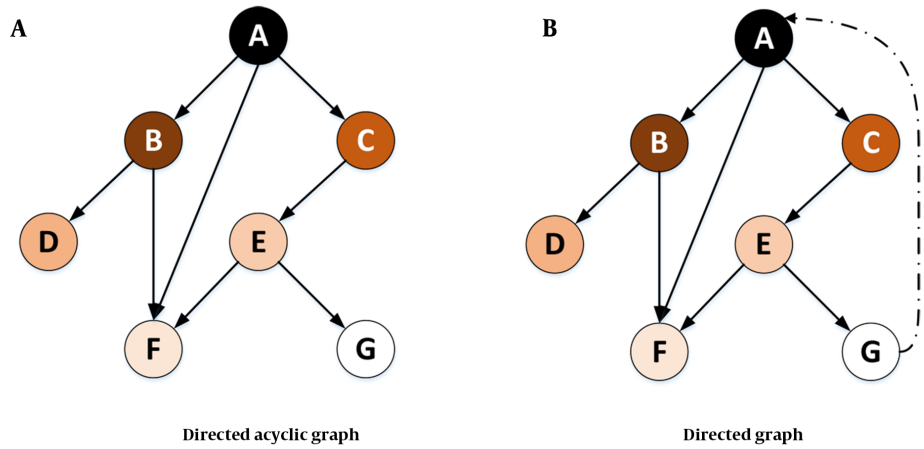

Figure 11. Two types of directed causal graphs: acyclic (A) and cyclic (B) models

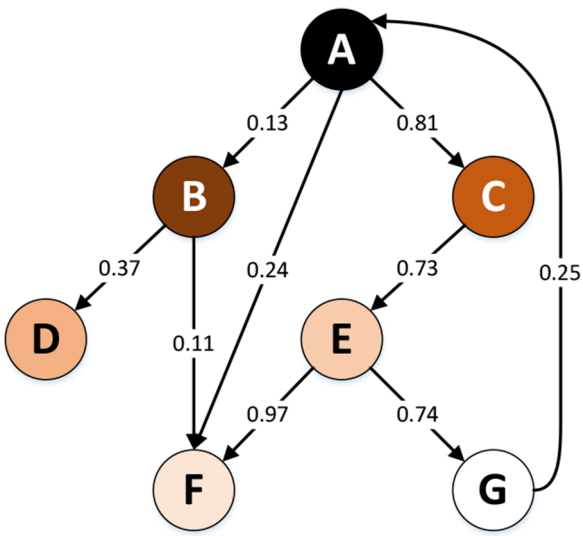

Figure 12. Directed weighted causal graphs

treatment approaches to control the disease and achieve more acceptable quality survivals and prognosis. In other words, in addition to identifying susceptible cases, the cause and effect of gene expressions in breast and CRC might be pursued, which might lead to further health status complications for each individual. Drawing the derivation tree of such a cause/effect relationship can reveal more implicit information. Such information can be helpful in the domain of precision medicine. To gain better in-depth insights, such relationships might be studied through network sciences (i.e., network medicine).

The clinical significance of this investigation is that machine learning algorithms could be used not only to improve diagnostic accuracy but also for identifying women at high risk of developing breast cancer and CRC, which could be prioritized for treatment (47). For supporting this idea, a study by Lux et al. it is shown that individuals with high genomic risk are among the substantial contributors to breast cancer treatment costs (53). It was studied that gene expression test was estimated to reduce costs versus standard care in Germany. Such cost-effectiveness seems to be true for both breast cancer and CRC (53, 54). However, it can be further studied and proved that utilizing computer-aided tools such as our proposed classifiers can reduce healthcare costs.

From a prognostic perspective, it is also shown that gene expression profiles for invasive early breast cancer have presented excellent prognostic capacities. This means that the gene expression profiles help for early diagnosis of the disease (55). Moreover, using machine learning-enabled tools such as the one provided in our study can reduce the need for clinical expertise to interpret the result and better and more accurately decide on a proper treatment plan.

One of the limitations of this study was the data size since two datasets consisting of 71 669 gene expressions were used. However, the authors devoted efforts to overcome such a limitation as much as possible, using the k-fold method.

5.1. Conclusions

The obtained findings revealed that in addition to achieving higher classification accuracy, the cause-effect

relationships as synthesis structures could be used to build Markov models in the clinical domain to analyze treatment paths and predict patient outcomes. According to this study, the cause-effect relationships are formed by gene expression data not necessarily depending on the genes triggering the disease but on genes indirectly affected by the disease. These sets of genes were obtained, using the proposed feature extraction method. The combination of such relationships with a probabilistic approach illuminates some implicit paths to find the best treatment options (i.e., best treatment paths), using graphic Markov models. Considering the gene expression data as the finite states of the Markov model provides a space, where any transitions of the states can help toward relevant predictions.

Footnotes

Authors' Contribution: Study concept and design: A. Moeini and M. Jafarpour; Acquisition of data: M. Jafarpour; Analysis and interpretation of data: M. Jafarpour and N. Maryami; Drafting of the manuscript: M. Jafarpour and N. Maryami; Critical revision of the manuscript for important intellectual content: A. Nahvijou and A. Mohammadian; Statistical analysis: M. Jafarpour and N. Maryami; Administrative, technical, and material support: A. Nahvijou; Study supervision: A. Moeini.

Conflict of Interests: There was no conflict of interest.

Ethical Approval: This study propose a method to classify cancer-prone and healthy cases under breast cancer and colorectal cancer (CRC) using machine learning methods and there is no need to obtain ethical approval code

Funding/Support: There was no funding support.

References

1. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*. 2015;**13**:8–17. [PubMed ID: 25750696]. [PubMed Central ID: PMC4348437]. https://doi.org/10.1016/j.csbj.2014.11.005.
2. Shafana ARF, Uwanthika GAI, Kartheeswaran T. Exploring the molecular subclasses and stage-specific genes of oral cancer: A bioinformatics analysis. *Cancer Treat Res Commun*. 2021;**27**:100320. [PubMed ID: 33545567]. https://doi.org/10.1016/j.ctarc.2021.100320.
3. Napoli S, Scuderi C, Gattuso G, Bella VD, Candido S, Basile MS, et al. Functional Roles of Matrix Metalloproteinases and Their Inhibitors in Melanoma. *Cells*. 2020;**9**(5). [PubMed ID: 32392801]. [PubMed Central ID: PMC7291303]. https://doi.org/10.3390/cells9051151.
4. Xiong Q, Jiao Y, Yang P, Liao Y, Gu X, Hu F, et al. The association study between CYP24A1 gene polymorphisms and risk of liver, lung and gastric cancer in a Chinese population. *Pathol Res Pract*. 2020;**216**(12):153237. [PubMed ID: 33065483]. https://doi.org/10.1016/j.prp.2020.153237.
5. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021;**71**(3):209–49. [PubMed ID: 33538338]. https://doi.org/10.3322/caac.21660.
6. Blumen H, Fitch K, Polkus V. Comparison of Treatment Costs for Breast Cancer, by Tumor Stage and Type of Service. *Am Health Drug Benefits*. 2016;**9**(1):23–32. [PubMed ID: 27066193]. [PubMed Central ID: PMC4822976].
7. Afkar A, Jalilian H, Pourreza A, Mir H, Sigaroudi AE, Heydari S. Cost analysis of breast cancer: a comparison between private and public hospitals in Iran. *BMC Health Serv Res*. 2021;**21**(1):219. [PubMed ID: 33706762]. [PubMed Central ID: PMC7953682]. https://doi.org/10.1186/s12913-021-06136-6.
8. Mariotto AB, Enewold L, Zhao J, Zeruto CA, Yabroff KR. Medical Care Costs Associated with Cancer Survivorship in the United States. *Cancer Epidemiol Biomarkers Prev*. 2020;**29**(7):1304–12. [PubMed ID: 32522832]. [PubMed Central ID: PMC9514601]. https://doi.org/10.1158/1055-9965.EPI-19-1534.
9. Blackadar CB. Historical review of the causes of cancer. *World J Clin Oncol*. 2016;**7**(1):54–86. [PubMed ID: 26862491]. [PubMed Central ID: PMC4734938]. https://doi.org/10.5306/wjco.v7.i1.54.
10. Shebli WTY, Alotibi MKH, Al-Raddadi RI, Al-Amri RJ, Fallatah EIY, Alhujaily AS, et al. Murine Double Minute 2 Gene (MDM2) rs937283A/G variant significantly increases the susceptibility to breast cancer in Saudi Women. *Saudi J Biol Sci*. 2021;**28**(4):2272–7. [PubMed ID: 33911942]. [PubMed Central ID: PMC8071807]. https://doi.org/10.1016/j.sjbs.2021.01.020.
11. Al-Khater KM, Almofty S, Ravinayagam V, Alrushaid N, Rehman S. Role of a metastatic suppressor gene KAI1/CD82 in the diagnosis and prognosis of breast cancer. *Saudi J Biol Sci*. 2021;**28**(6):3391–8. [PubMed ID: 34121877]. [PubMed Central ID: PMC8176039]. https://doi.org/10.1016/j.sjbs.2021.03.001.
12. Yang Y, Leonard M, Luo Z, Yeo S, Bick G, Hao M, et al. Functional cooperation between co-amplified genes promotes aggressive phenotypes of HER2-positive breast cancer. *Cell Rep*. 2021;**34**(10):108822. [PubMed ID: 33691110]. [PubMed Central ID: PMC8050805]. https://doi.org/10.1016/j.celrep.2021.108822.
13. Injinari N, Amini-Farsani Z, Yadollahi-Farsani M, Teimori H. Apoptotic effects of valproic acid on miR-34a, miR-520h and HDAC1 gene in breast cancer. *Life Sci*. 2021;**269**:119027. [PubMed ID: 33453248]. https://doi.org/10.1016/j.lfs.2021.119027.
14. Haddad S. CYP1A1* 2 A gene polymorphism frequency in Syrian breast cancer patients. *Meta Gene*. 2021;**28**:100861. https://doi.org/10.1016/j.mgene.2021.100861.
15. Beesley J, Sivakumaran H, Moradi Marjaneh M, Shi W, Hillman KM, Kaufmann S, et al. eQTL Colocalization Analyses Identify NTN4 as a Candidate Breast Cancer Risk Gene. *Am J Hum Genet*. 2020;**107**(4):778–87. [PubMed ID: 32871102]. [PubMed Central ID: PMC7536644]. https://doi.org/10.1016/j.ajhg.2020.08.006.
16. Du JX, Chen C, Luo YH, Cai JL, Cai CZ, Xu J, et al. Establishment and validation of a novel autophagy-related gene signature for patients with breast cancer. *Gene*. 2020;**762**:144974. [PubMed ID: 32707305]. https://doi.org/10.1016/j.gene.2020.144974.
17. Mohammadian H, Sharifi R, Rezanezhad Amirdehi S, Taheri E, Babazadeh Bedoustani A. Matrix metalloproteinase MMP1 and MMP9 genes expression in breast cancer tissue. *Gene Reports*. 2020;**21**:100906. https://doi.org/10.1016/j.genrep.2020.100906.
18. Zhang LZ, Qi WH, Zhao G, Liu LX, Xue H, Hu WX, et al. Correlation between PTEN and P62 gene expression in rat colorectal cancer cell. *Saudi J Biol Sci*. 2019;**26**(8):1986–90. [PubMed ID: 31885487]. [PubMed Central ID: PMC6921302]. https://doi.org/10.1016/j.sjbs.2019.08.006.
19. Irham LM, Wong HS, Chou WH, Adikusuma W, Mugiyanto E, Huang WC, et al. Integration of genetic variants and gene network for drug repurposing in colorectal cancer. *Pharmacol Res*. 2020;**161**:105203. [PubMed ID: 32950641]. https://doi.org/10.1016/j.phrs.2020.105203.

20. Chang Z, Liu X, Zhao W, Xu Y. Identification and Characterization of the Copy Number Dosage-Sensitive Genes in Colorectal Cancer. *Mol Ther Methods Clin Dev*. 2020;**18**:501–10. [PubMed ID: 32775488]. [PubMed Central ID: PMC7390836]. https://doi.org/10.1016/j.omtm.2020.06.020.
21. Caiado H, Conceicao N, Tiago D, Marreiros A, Vicente S, Enriquez JL, et al. Data on the evaluation of FGF2 gene expression in Colorectal Cancer. *Data Brief*. 2020;**31**:105765. [PubMed ID: 32551343]. [PubMed Central ID: PMC7289741]. https://doi.org/10.1016/j.dib.2020.105765.
22. Khalaj-Kondori M, Hosseinnejad M, Hosseinzadeh A, Behroz Sharif S, Hashemzadeh S. Aberrant hypermethylation of OGDHL gene promoter in sporadic colorectal cancer. *Curr Probl Cancer*. 2020;**44**(1):100471. [PubMed ID: 30904169]. https://doi.org/10.1016/j.currproblcancer.2019.03.001.
23. Sadeghi H, Nazemalhosseini-Mojarad E, Yassaee VR, Savabkar S, Ghasemian M, Aghdaei HA, et al. Could CYP24A1 promoter methylation status affect the gene expression in the colorectal cancer patients? *Meta Gene*. 2020;**24**:100656. https://doi.org/10.1016/j.mgene.2020.100656.
24. Battagin AS, Bertuzzo CS, Carvalho PO, Ortega MM, Marson FAL. Single nucleotide variants c.-13G –> C (rs17429833) and c.108C –> T (rs72466472) in the CLDN1 gene and increased risk for familial colorectal cancer. *Gene*. 2021;**768**:145304. [PubMed ID: 33186612]. https://doi.org/10.1016/j.gene.2020.145304.
25. Mo HY, An CH, Choi EJ, Yoo NJ, Lee SH. Somatic mutation and loss of expression of a candidate tumor suppressor gene TET3 in gastric and colorectal cancers. *Pathol Res Pract*. 2020;**216**(3):152759. [PubMed ID: 31859118]. https://doi.org/10.1016/j.prp.2019.152759.
26. Bai JP, Alekseyenko AV, Statnikov A, Wang IM, Wong PH. Strategic applications of gene expression: from drug discovery/development to bedside. *AAPS J*. 2013;**15**(2):427–37. [PubMed ID: 23319288]. [PubMed Central ID: PMC3675744]. https://doi.org/10.1208/s12248-012-9447-1.
27. Burska AN, Roget K, Blits M, Soto Gomez L, van de Loo F, Hazelwood LD, et al. Gene expression analysis in RA: towards personalized medicine. *Pharmacogenomics J*. 2014;**14**(2):93–106. [PubMed ID: 24589910]. [PubMed Central ID: PMC3992869]. https://doi.org/10.1038/tpj.2013.48.
28. Grant GR, Manduchi E, Stoeckert CJ. Analysis and management of microarray gene expression data. *Curr Protoc Mol Biol*. 2007;**Chapter 19**:Unit 19 6. [PubMed ID: 18265395]. https://doi.org/10.1002/0471142727.mb1906s77.
29. Werner T. Bioinformatics applications for pathway analysis of microarray data. *Curr Opin Biotechnol*. 2008;**19**(1):50–4. [PubMed ID: 18207385]. https://doi.org/10.1016/j.copbio.2007.11.005.
30. Bashiri A, Ghazisaeedi M, Safdari R, Shahmoradi L, Ehtesham H. Improving the Prediction of Survival in Cancer Patients by Using Machine Learning Techniques: Experience of Gene Expression Data: A Narrative Review. *Iran J Public Health*. 2017;**46**(2):165–72. [PubMed ID: 28451550]. [PubMed Central ID: PMC5402773].
31. Ozer ME, Sarica PO, Arga KY. New Machine Learning Applications to Accelerate Personalized Medicine in Breast Cancer: Rise of the Support Vector Machines. *OMICS*. 2020;**24**(5):241–6. [PubMed ID: 32228365]. https://doi.org/10.1089/omi.2020.0001.
32. Zhi J, Sun J, Wang Z, Ding W. Support vector machine classifier for prediction of the metastasis of colorectal cancer. *Int J Mol Med*. 2018;**41**(3):1419–26. [PubMed ID: 29328363]. [PubMed Central ID: PMC5819940]. https://doi.org/10.3892/ijmm.2018.3359.
33. de Ronde JJ, Bonder MJ, Lips EH, Rodenhuis S, Wessels LF. Breast cancer subtype specific classifiers of response to neoadjuvant chemotherapy do not outperform classifiers trained on all subtypes. *PLoS One*. 2014;**9**(2). e88551. [PubMed ID: 24558399]. [PubMed Central ID: PMC3928239]. https://doi.org/10.1371/journal.pone.0088551.
34. Smolander J, Dehmer M, Emmert-Streib F. Comparing deep belief networks with support vector machines for classifying gene expression data from complex disorders. *FEBS Open Bio*. 2019;**9**(7):1232–48. [PubMed ID: 31074948]. [PubMed Central ID: PMC6609581]. https://doi.org/10.1002/2211-5463.12652.
35. Wang H, Xiong J, Yao Z, Lin M, Ren J. Research Survey on Support Vector Machine. *10th EAI International Conference on Mobile Multimedia Communications*. EAI; 2017.
36. Aydadenta H, Adiwijaya A. A clustering approach for feature selection in microarray data classification using random forest. *J Inform Proc Sys*. 2018;**14**(5):1167–75. https://doi.org/10.3745/JIPS.04.0087.
37. Kuhn M, Johnson K. *Applied Predictive Modeling*. Springer; 2013.
38. Phuong TM, Lin Z, Altman RB. Choosing SNPs using feature selection. *2005 IEEE Computational Systems Bioinformatics Conference (CSB'05)*. Stanford, USA. IEEE; 2005. p. 301–9.
39. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;**30**(1):207–10. [PubMed ID: 11752295]. [PubMed Central ID: PMC99122]. https://doi.org/10.1093/nar/30.1.207.
40. Torres-Barrán A, Alaíz CM, Dorronsoro JR. Faster SVM training via conjugate SMO. *Pattern Recognition*. 2021;**111**:107644. https://doi.org/10.1016/j.patcog.2020.107644.
41. Chiu H, Li TS, Kuo P. Breast Cancer–Detection System Using PCA, Multilayer Perceptron, Transfer Learning, and Support Vector Machine. *IEEE Access*. 2020;**8**:204309–24. https://doi.org/10.1109/access.2020.3036912.
42. Xu X, Zhang Y, Zou L, Wang M, Li A. A gene signature for breast cancer prognosis using support vector machine. *2012 5th International Conference on BioMedical Engineering and Informatics*. Chongqing, China. IEEE; 2012. p. 928–31.
43. Egwom OJ, Hassan M, Tanimu JJ, Hamada M, Ogar OM. An LDA-SVM Machine Learning Model for Breast Cancer Classification. *BioMedInformatics*. 2022;**2**(3):345–58. https://doi.org/10.3390/biomedinformatics2030022.
44. Naji MA, Filali SE, Aarika K, Benlahmar EH, Abdelouhahid RA, Debauche O. Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. *Procedia Comput Sci*. 2021;**191**:487–92. https://doi.org/10.1016/j.procs.2021.07.062.
45. Aljuaid H, Alturki N, Alsubaie N, Cavallaro L, Liotta A. Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning. *Comput Methods Programs Biomed*. 2022;**223**:106951. [PubMed ID: 35767911]. https://doi.org/10.1016/j.cmpb.2022.106951.
46. Arooj S, Atta Ur R, Zubair M, Khan MF, Alissa K, Khan MA, et al. Breast Cancer Detection and Classification Empowered With Transfer Learning. *Front Public Health*. 2022;**10**:924432. [PubMed ID: 35859776]. [PubMed Central ID: PMC9289190]. https://doi.org/10.3389/fpubh.2022.924432.
47. Wu J, Hicks C. Breast Cancer Type Classification Using Machine Learning. *J Pers Med*. 2021;**11**(2). [PubMed ID: 33498339]. [PubMed Central ID: PMC7909418]. https://doi.org/10.3390/jpm11020061.
48. Fakoor R, Ladhak F, Nazi A, Huber M. Using deep learning to enhance cancer diagnosis and classification. *Proceedings of the international conference on machine learning*. New York, USA. ACM; 2013. p. 3937–49.
49. Sherr CJ. Principles of tumor suppression. *Cell*. 2004;**116**(2):235–46. [PubMed ID: 14744434]. https://doi.org/10.1016/s0092-8674(03)01075-4.
50. Kumar A, Gosipatala SB, Pandey A, Singh P. Prognostic Relevance of SFRP1 Gene Promoter Methylation in Colorectal Carcinoma. *Asian Pac J Cancer Prev*. 2019;**20**(5):1571–7. [PubMed ID: 31128064]. [PubMed Central ID: PMC6857878]. https://doi.org/10.31557/APJCP.2019.20.5.1571.
51. Chung SJ, Nagaraju GP, Nagalingam A, Muniraj N, Kuppusamy P, Walker A, et al. ADIPOQ/adiponectin induces cytotoxic autophagy in breast cancer cells through STK11/LKB1-mediated activation of the AMPK-ULK1 axis. *Autophagy*. 2017;**13**(8):1386–403. [PubMed ID: 28696138]. [PubMed Central ID: PMC5584870]. https://doi.org/10.1080/15548627.2017.1332565.
52. Piccininni M, Konigorski S, Rohmann JL, Kurth T. Directed acyclic graphs and causal thinking in clinical risk prediction modeling. *BMC*

Med Res Methodol. 2020;**20**(1):179. [PubMed ID: 32615926]. [PubMed Central ID: PMC7331263]. https://doi.org/10.1186/s12874-020-01058-z.

53. Lux MP, Nabieva N, Hildebrandt T, Rebscher H, Kummel S, Blohmer JU, et al. Budget impact analysis of gene expression tests to aid therapy decisions for breast cancer patients in Germany. *Breast*. 2018;**37**:89–98. [PubMed ID: 29128582]. https://doi.org/10.1016/j.breast.2017.11.002.
54. Ramdzan AR, Manaf MRA, Aizuddin AN, Latiff ZA, Teik KW, Ch'ng GS, et al. Cost-Effectiveness of Colorectal Cancer Genetic Testing. *Int J Environ Res Public Health*. 2021;**18**(16). [PubMed ID: 34444091]. [PubMed Central ID: PMC8394708]. https://doi.org/10.3390/ijerph18168330.
55. Blok EJ, Bastiaannet E, van den Hout WB, Liefers GJ, Smit V, Kroep JR, et al. Systematic review of the clinical and economic value of gene expression profiles for invasive early breast cancer available in Europe. *Cancer Treat Rev*. 2018;**62**:74–90. [PubMed ID: 29175678]. https://doi.org/10.1016/j.ctrv.2017.10.012.