Cell J
Cell Journal(Yakhteh)
Homepage: https://www.celljournal.org/
Original Article
Vol 25, No 5, May 2023, Pages: 347-353

The Performance Evaluation of The Random Forest Algorithm for A Gene Selection in Identifying Genes Associated with Resectable Pancreatic Cancer in Microarray Dataset: A Retrospective Study

Niloofar Rabiei, Ph.D.[1], Ali Reza Soltanian, Ph.D.[2]*, Maryam Farhadian, Ph.D.[3], Fatemeh Bahreini, Ph.D.[4]

1. Department of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran
2. Modeling of Noncommunicable Diseases Research Center, Hamadan University of Medical Sciences, Hamadan, Iran
3. Research Center for Health Sciences, Hamadan University of Medical Sciences, Hamadan, Iran
4. Department of Molecular Medicine and Genetics, School of Medicine, Hamadan University of Medical Sciences, Hamadan, Iran

Abstract

Objective: In microarray datasets, hundreds and thousands of genes are measured in a small number of samples, and sometimes due to problems that occur during the experiment, the expression value of some genes is recorded as missing. It is a difficult task to determine the genes that cause disease or cancer from a large number of genes. This study aimed to find effective genes in pancreatic cancer (PC). First, the K-nearest neighbor (KNN) imputation method was used to solve the problem of missing values (MVs) of gene expression. Then, the random forest algorithm was used to identify the genes associated with PC.

Materials and Methods: In this retrospective study, 24 samples from the GSE14245 dataset were examined. Twelve samples were from patients with PC, and 12 samples were from healthy control. After preprocessing and applying the fold-change technique, 29482 genes were used. We used the KNN imputation method to impute when a particular gene had MVs. Then, the genes most strongly associated with PC were selected using the random forest algorithm. We classified the dataset using support vector machine (SVM) and naïve bayes (NB) classifiers, and F-score and Jaccard indices were reported.

Results: Out of the 29482 genes, 1185 genes with fold-changes greater than 3 were selected. After selecting the most associated genes, 21 genes with the most important value were identified. *S100P* and *GPX3* had the highest and lowest importance values, respectively. The F-score and Jaccard value of the SVM and NB classifiers were 95.5, 93, 92, and 92 percent, respectively.

Conclusion: This study is based on the application of the fold change technique, imputation method, and random forest algorithm and could find the most associated genes that were not identified in many studies. We therefore suggest researchers use the random forest algorithm to detect the related genes within the disease of interest.

Keywords: Classification, Microarray Analysis, Neoplasms, Pancreas

Citation: **Rabiei N, Soltanian AR, Farhadian M, Bahreini F. The performance evaluation of the random forest algorithm for a gene selection in identifying genes associated with resectable pancreatic cancer in microarray dataset: a retrospective study. 2023; 25(5): 347-353. doi: 0.22074/CELLJ.2023.1971852.1156**

Introduction

One of the deadliest diseases and malignancies is PC (1). According to the Global Cancer Observatory (GLOBOCAN) 2020 report, 495773 new patients were diagnosed with PC in 2020 (https://gcoiarcfr/today/home). The results of PC trends in 48 countries showed an increase in PC incidence in women and men in 17 and 14 countries, respectively (2).

According to the American Cancer Society, 60430 and 48220 new PC patients and deaths were reported in the United States in 2021 (3). Some studies have shown that the incidence and mortality rates are higher in men in Iran (4). The incidence rate in females is 2.26 and 1.24 percent in North Khorasan and Chaharmahal and Bakhtiari provinces, respectively. The incidence rate in males is 1.17 and 1.12 percent in Semnan and Markazi provinces, respectively (5).

Early detection or diagnosis of PC requires efficient molecular and screening methods. One of these methods is the identification of the gene set associated with this disease. Microarray is the most efficient technique to achieve this goal. Microarrays are a powerful technology that can monitor more than thousands of genes simultaneously on a single chip. They are usually represented in the form of a matrix, with rows corresponding to genes and columns corresponding to different conditions (6).

Since there are thousands of genes in microarray data, identifying and prioritizing genes significantly associated

Received: 02/December/2022, Revised: 14/January/2023, Accepted: 15/February/2023
***Corresponding Address: P.O.BOX: 6517838736, Modeling of Noncommunicable Diseases Research Center, Hamadan University of Medical Sciences, Hamadan, Iran**
Email: soltanian@umsha.ac.ir

Royan Institute
Cell Journal (Yakhteh)

with the disease can play an important role in the diagnosis process and in reducing the mortality rate. One of the techniques to select genes related to the disease is random forest, which is very popular in the field of gene/feature selection (7).

To improve the performance of an analysis, e.g., in identifying the most important genes, clustering, and classification analysis, the dataset must be complete and not contain missing values (MVs). Like many other datasets, gene expression data often contain MVs ranging from 5 to 60 percent (8, 9). Poor hybridization, insufficient resolution or corruption to the image, and contamination from dust or scratches on the slide can lead to the creation of MVs. However, methods such as classification and clustering require a complete dataset as input. So, in such cases, dealing with MVs is a very important step. The simplest way is to discard the observations with MVs or impute MVs with imputation methods, e.g., the KNN imputation method (10, 11).

This study aimed to select the most important genes associated with resectable PC. For this purpose, MVs were imputed using the best imputation method between mean, median, Multiple Imputation with Denoising Autoencoders (MIDAS), and KNN imputation methods Then, the random forest algorithm was used to identify the associated genes. After the imputation and selection of the most strongly associated genes, a bi-clustering method was applied, and a heat map was generated based on all genes and the selected genes. The precision value of the support vector machine (SVM) classifier was also reported. To better understand the effects of gene selection, another strategy called the complete case strategy was also considered. In this strategy, all genes with MVs are excluded from the dataset. Then, gene selection is performed using the random forest algorithm. The resulting heatmap is examined. The strength and novelty of this research is the use of the most effective machine learning methods, random forest, to select effective genes for the disease, and it is not only limited to fold change values to select effective genes. In addition, this study paid special attention to imputation as one of the most important steps in data preprocessing. After evaluating the most common imputation methods, the best method is selected to ensure that the subsequent analyses have the least bias and error.

Materials and Methods

Dataset

In this retrospective study, we used data from Zhang et al. (12) with accession number GSE14245, which is available in the GEO database. This dataset contains 12 PC and 12 healthy control samples.

This study was approved by the Ethics Committee of Hamadan University of Medical Sciences (IR.UMSHA.REC.1400.212).

Imputation of missing values

If the dataset contains MVs, they are imputed first. The quality of the inference depends on the rate of missing genes. To compare methods for imputing missing data, some studies consider genes with less than 1 percent, between 1 - 5, or 1 - 20 percent MVs. However, there is no accepted threshold for the rate of missing genes to determine whether imputation should be performed (13, 14). Some studies reported that less than 1% of MVs are considered nonsignificant, 1-5% are controllable, and more than 15% affect prediction or interpretation too much (15). In most studies, the rate of missing is high, and repeating the experiment is not feasible due to the high cost or time constraints. To perform a more accurate study, the authors of this study decided to consider genes with a missing rate of less than 25% and remove the others. To find the best imputation method between the KNN, mean, median, and MIDAS imputation methods, the MVs in the genes are first imputed according to the imputation method (KNN, mean, median, and MIDAS). The imputed data are classified using the logistic classifier. The best imputation method is the one that leads to the highest value of classification accuracy. Each imputation method has been explained in the following sections:

K-nearest neighbor imputation method

Consider the matrix with n samples and p genes. The elements of G are represented by , where i=1,…,n and j=1,…,p. In the KNN imputation method, to impute MVs, we first calculate the Euclidean distance between i=1,…,n , j=1,…,p, whose value is missing, and all genes without MVs. The K genes with the smallest distance are selected. The MVs in are imputed by averaging the expression values of K-selected genes. The parameter K should be chosen experimentally. We chose the best K value between 5 and 500 in 5 steps.

Mean and median imputation method

In the mean imputation method, the MVs in genes are imputed by the average of all samples in that gene that are not missing. In the median imputation method, the MVs in genes are imputed by the median of all samples in that gene that are not missing.

Multiple imputation with denoising autoencoders imputation method

The MIDAS method treats the MVs as a part of the corrupted data and attempts to impute the MVs using a model trained to minimize the specific error. MIDAS can capture the complex relationship between genes. To reduce overfitting during imputation, the dropout technique was used (16).

Fold-change calculation

After imputation, genes whose fold-change was greater than the specific α-value were selected for the next stage of analysis. Fold-change is defined as equation 1,

$$\text{Fold-change} = 2^{\left|\log_2(\text{average(case)}) - \log_2(\text{average(control)})\right|}, \quad \text{[Equation 1]}$$

where the average (case) and average (control) are the expression values of genes in the case and control groups, respectively.

Random forest algorithm

Gene selection using the random forest algorithm was performed while the forest was growing. First, a random subset of genes is taken from the dataset, and n random trees are created from each subset. Then, at the end of the process, all the created trees are combined to extract the corresponding results. The common index to measure the importance of genes is the Gini index for importance/impurity. In each tree, the sum of the Gini index reduction is calculated over all nodes for the specific gene used for partitioning.

If the dataset of the two classes labeled 0 and 1 in node τ wants to be split, the Gini index at node τ, Gini (τ), is calculated as in equation 2,

$$\text{Gini}(\tau)=1-p_1^2-p_0^2, \text{ [Equation 2]}$$

where p_i $i=0,1$ the probability that sample belongs to class i.

The impurity reduction for splitting the samples between and (right and left node) is described in Equation 3,

$$\Delta \text{ Gini (G)} = \text{Gini}(\tau) - p_L \text{ Gini}(\tau_L) - p_R \text{ gini}(\tau_R), \text{ [Equation 3]}$$

where , is the probability of sample of node i. Equation 2 is calculated for all available thresholds, θ's, over all genes at node . The gene with threshold θ that maximizes is determined. After extracting the optimal for all nodes and trees (T), the Gini importance, , is calculated for a given gene i.e., G, according to equation 4.

$$I_{\text{Gini}}^{(G)}=\sum_{(\text{all nodes})} \sum_{(\text{all trees})} \Delta\text{Gini}_\theta (\tau,T), \text{ [Equation 4]}$$

shows the discrimination degree of G for diagnosis between two groups (case and control) and also indicates how many times a G was selected in the splitting process (8). Figure 1 shows the simple tree with four genes, . The values are the thresholds that lead to the highest values.

In this study, imputation was performed using the impute package in R software version 4.2.1, and gene selection using the random forest algorithm was performed using the sklearn module in Python 3.9 software.

Classification and classification index

The naïve bayes (NB) classifier is a probabilistic classification mechanism for finding the best result in classification problems. This method assumes that variables are independent for a given class, and can be represented as

$$P(X \mid C) = \prod_{(i=1)}^{n} P(X_i \mid C)$$

where $X=(X_1,\ldots,X_{n)}$ is a variable vecor and C is a class label. SVM is a learning method used for both classification and regression. The idea behind SVM is that all data are first plotted in an n-dimensional space. Classification is then used to find a hyperplane that can separate the data based on class membership. We used Python software version 3.9 to analyze the dataset. In binary classification, there are four possible outcomes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). TP, TN, FP, and FN are defined as a correct result, correct absence of result, unexpected result, and missing result, respectively. Different combinations of outcomes were used in each Jaccard and F score index. The Jaccard index or Jaccard similarity coefficient measures the similarity of sample groups and is defined as TP/(TP+FP+FN). F-score or F1-measure defined as (2×(precision×recall))/(precision+recall). Recall and precision defined as TP/(TP+FN) and TP/(TP+FP), respectively. These indices reach their best value at 1 and worst value at 0.

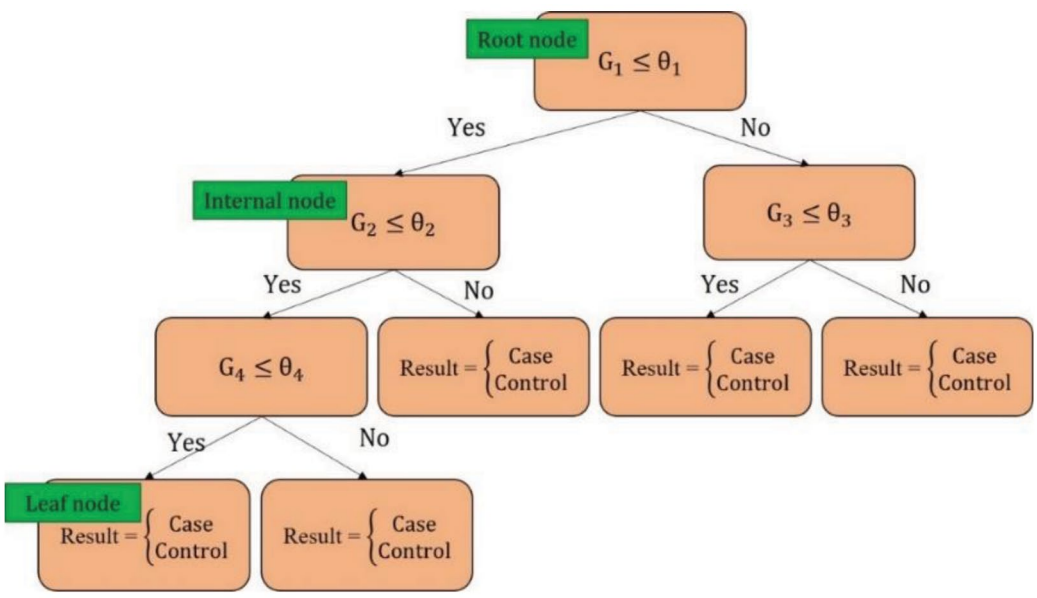

Fig.1: The tree with four gene, three internal node, and five leaf nodes.

Results

In this study, the pattern of the MVs is non-monotone or general, and the mechanism of MVs was complete at random. Any gene with MVs greater than 25 percent was discarded and removed from the dataset. Out of the 29482 remaining genes, the MVs were imputed by the KNN imputation method with K value 5. After imputation, the genes with fold-change greater than 3 were selected. Different cut points were considered in different studies. For example, cut points between 1.8 and 3, 1.5, 2, or 4 were also considered. In this study, cut points 2, 3, and 4 were considered, and the best result was obtained with cut-point (17, 18). Among 29482 genes, 1185 genes were selected. The random forest algorithm was used to select the genes most strongly associated with PC. The genes whose I_{Gini} (G) was greater than or equal to 0.008 were selected. As mentioned earlier, the random forest algorithm first calculates the I_{Gini} (G) for all genes and then extracts the average, median, or percentile of these values. Genes with I_{Gini} (G) values greater than the average, median, or percentile are selected as important genes. The information on genes is provided in Table 1.

Table 1: A gene set associated with pancreatic cancer based on the random forest algorithm

Gene symbol	Gene ID	Gene title/Description	Chromosome	Cytoband	Missing rate (%)	Fold-change*	I_{Gini} (G)**
S100P	6286	S100 calcium binding protein P	4	p16.1	4	11.169	0.033
CDC14B	8555	Cell division cycle 14B	9	q22.32-q22.33	20	8.105	0.010
CD7	924	CD7 molecule	17	q25.3	20	13.133	0.021
CDH4	1002	Cadherin 4	20	q13.33	0	8.924	0.018
ZMIZ2	83637	Zinc finger MIZ-type containing 2	7	p13	0	7.444	0.017
LRRK1	79705	Leucine rich repeat kinase 1	15	q26.3	0	19.181	0.016
PERP	64065	PERP, TP53 apoptosis effector	6	q23.3	4	8.422	0.015
LILRA2	11027	Leukocyte immunoglobulin like receptor A2	19	q13.42	12	9.443	0.014
ENG	2022	Endoglin	9	q34.11	0	6.198	0.013
APOH	350	Apolipoprotein H	17	q24.2	16	3.018	0.012
ITGA2B	3674	Integrin subunit alpha 2b	17	q21.31	0	3.986	0.010
ACRV1	56	Acrosomal vesicle protein 1	11	q24.2	4	5.525	0.010
GALNT6	11226	Polypeptide N-acetylgalactosaminyltransferase 6	12	q13.13	20	7.266	0.009
DTX3	196403	Deltex E3 ubiquitin ligase 3	12	q13.3	0	4.198	0.009
DDX3X	1654	DEAD-box helicase 3, X-linked	X	p11.4	0	4.497	0.009
FTH1	2495	Ferritin heavy chain 1	11	q12.3	0	24.540	0.009
EME2	197342	Essential meiotic structure-specific endonuclease subunit 2	16	p13.3	4	3.090	0.008
HOOK1	51361	Hook microtubule tethering protein 1	1	p32.1	12	3.128	0.008
KRAS	3845	KRAS proto-oncogene, GTPase/ Kirsten rat sarcoma viral oncogene homolog	12	p12.1	20	4.506	0.008
SULF2	55959	Sulfatase 2	20	q13.12	20	3.701	0.008
GPX3	2878	Glutathione peroxidase 3	5	q33.1	4	3.659	0.008

*; Fold-change calculated by equation 1 and **; Random forest algorithm.

The heatmap of the dataset based on 1185 genes (genes with a fold-change greater than 3) is shown in Figure 2. Samples are shown on the horizontal axis and genes on the vertical axis. The vertical axis is the result of gene clustering using the Euclidean distance criterion and the hierarchical single linkage clustering algorithm. As can be seen, the separation between PC (case) and healthy control (control) groups was not perfect. A regular pattern could not be detected, indicating that the discriminatory power of 1185 genes was not satisfactory.

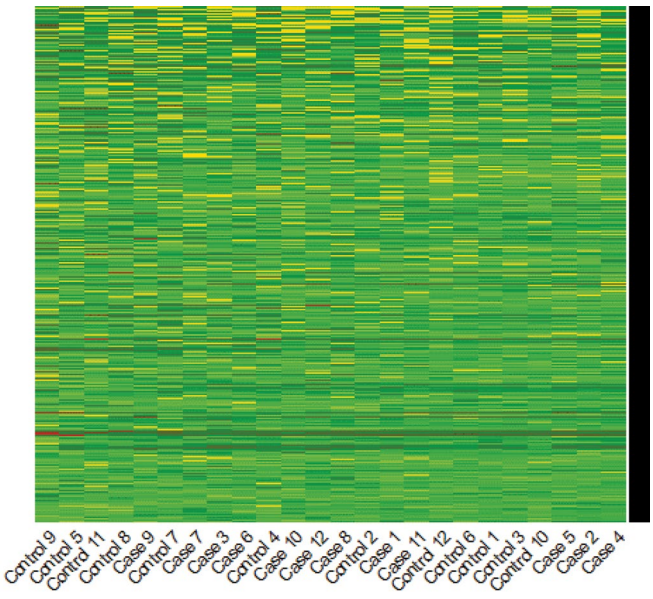

Fig.2: Heatmap derived from bi-clustering based on genes with the fold-change greater than 3 (n=1185). The colors on the heatmap show the expression levels from yellow (lowest) to red (highest).

Figure 3 shows the heatmap of the dataset based on 22 genes selected by the random forest algorithm. As in Figure 2, samples are shown on the horizontal axis and genes are shown on the vertical axis. The vertical axis shows the result of gene clustering using the Euclidean distance criterion and Ward.D linkage hierarchical clustering algorithm. As shown, the genes selected by random forest were able to reveal the distinction between the groups.

Using the complete case strategy, 10 genes were selected from 8977 fully expressed genes with fold-change greater than 3 and the greater than or equal to 0.024. Figure 4 shows the heatmap of the dataset based on these selected genes using the random forest algorithm. As shown, no regular pattern was detected and the separation between the two groups of PC (case) and healthy control (control) was not satisfactory.

The classification results for the complete case dataset (dataset A), the dataset with 1185 genes extracted based on fold-change greater than 3 (dataset B), and the dataset with 21 genes extracted based on fold-change greater than 3 and selected using the random forest algorithm (dataset C) are shown in Table 2. As shown, dataset C led to the best result with Jaccard and F-score values of 83 and 93 percent respectively, for the NB classifier. The Jaccard and F-score values of the SVM classifier based on dataset C were 92 and 96 percent, respectively.

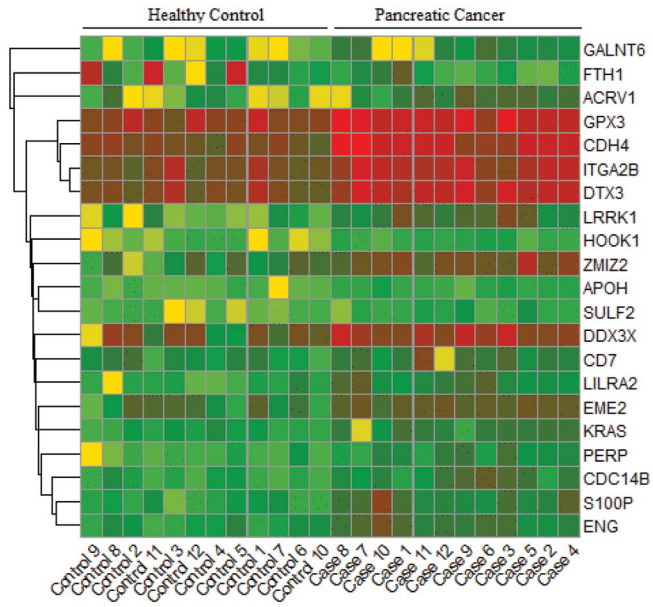

Fig.3: Heatmap derived from bi-clustering based on genes with fold-change greater than 3 and selected by random forest algorithm (n=21). The colors on the heatmap show the expression levels from yellow (lowest) to red (highest).

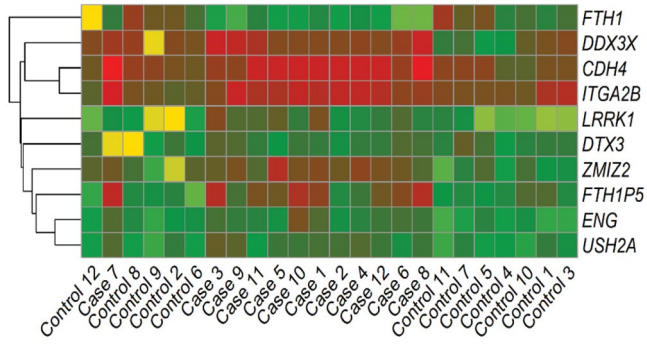

Fig.4: Heatmap derived from bi-clustering based on genes with fold-change greater than 3 and selected by random forest algorithm (n=10) from the complete case dataset. The colors on the heatmap show the expression levels from yellow (lowest) to red (highest).

Table 2: JThe percent of the Jaccard and F-score values of NB and SVM classifiers for datasets A, B, and C

Classifier	Dataset A		Dataset B		Dataset C	
	Jaccard	F-score	Jaccard	F-score	Jaccard	F-score
NB	53	67	81	90	92	93
SVM	78	87	83	90	92	95.5

NB; Naïve Bayes and SVM; Support vector machine. The complete case dataset (dataset A), The dataset with 1185 genes (dataset B), and the dataset with 21 genes (dataset C).

Discussion

In this study, high dimensionality has been discussed as one of the major challenges in the analysis of microarray datasets. Microarray datasets contain hundreds and thousands of genes, but not all of them are associated with the disease. Therefore, it is very important to find the most associated genes, and the use of an appropriate method is essential for this task. We used the random forest algorithm for selecting the most associated genes. The random forest has several properties that make it ideal for the analysis of microarray datasets. It can be used when there are too many genes but few samples. On the other hand, it works well when faced with a high-dimensional dataset. Even in the presence of noise, its prediction performance was good. It provides the Gini importance value for each gene. Although we do not have a categorical variable (i.e., sample sex) in this study, it could be treated as both a categorical and continuous variable (19). It should also be noted that the microarray dataset usually had MVs, so before any analysis of the available dataset, the MVs were imputed. To determine the best imputation method, the KNN method was chosen among the different methods. Several studies have reported that the KNN performed the best in different situations (11, 20). As mentioned earlier, one strategy for MVs is to remove genes with MVs and analyze the dataset with the complete genes. In this study, after removing MVs, the random forest algorithm was used, and the results showed that not only were the selected genes unable to identify the difference between the healthy control group and the PC group, but also some important genes such as *S100P, CDC14B,* and *CD7* were removed from the dataset due to missingness.

The SVM and NB classifiers were used for further analysis. The classification result for the complete case dataset (dataset A), the dataset with 1185 genes extracted based on fold-change greater than 3 (dataset B), and the dataset with 21 genes extracted based on fold-change greater than 3 and selected by the random forest algorithm (dataset C) were reported. Using the SVM classifier, the minimum F-score and Jaccard value belong to dataset A, and the maximum F-score and Jaccard value belong to dataset C. The results also show that the SVM classifier has better performance than the NB classifier.

In this study, 21 genes associated with PC were selected from 1185 genes with fold-change greater than 3. Among the genes selected in this study, *S100P*, *CD7*, *CDH4*, *ZMIZ2*, *LRRK1*, *LILRA2*, *ENG*, *ITGA2B*, *DTX3*, *FTH1*, *KRAS*, *CDC14B*, *ACRV1*, *DDX3X*, and *GPX3* were also identified in the Zhang et al. (12) study. They used quantitative polymerase chain reaction (PCR) to investigate the genes associated with PC. Some genes were identified in the study by Zhang et al. (12) but not in our study (i.e., *ASH2L*, *CABLES1*, and *CDKL3*). *ASH2L* had a fold-change greater than 3 (fold-change = 9.009), but its importance value was less than 0.008 (importance value=0.003). *CABLES1* also had a fold-change greater than 3 (fold-change=6.061), but its importance value was less than 0.008 (importance value=0.004). *CDKL3* had a fold-change of less than 3 (fold-change=1.500).

S100P has been introduced as a prognostic gene in PC in some studies (21, 22). In the study by Bardeesy and DePinho's (23) and the study by Kamisawa et al. (24), *KRAS* was introduced as a prognostic gene in PC. *PERP* was also the prognostic marker in PC (25). There is also some evidence of an association between *PERP* and PC on the Human Protein Atlas website (https://www.proteinatlas.org/). Park et al. (26) mentioned in their study that *FTH1* is associated with PC. Fujiwara et al. (27) also mentioned in their study that *ENG* is one of the prognostic genes in PC. Ouyang et al. (28) stated that *DTX3* is a prognostic gene in PC. Li et al. (29) stated that *CDH4* is associated with PC. Our results showed that *APOH* was a prognostic gene in PC. Kuwae et al. (30) stated in their study that *APOH* is one of the prognostic factors in human pancreatic ductal adenocarcinomas. Hao et al. (31) mentioned that *DDX3X* was associated with PC. Like our study, Alhasan et al. (32) also showed that *SULF2* was associated with pancreatic ductal adenocarcinoma. Tarhan et al. (33) stated that *GALNT6* was associated with PC, which was in line with our findings. In our study, *HOOK1* was identified as one of the genes associated with PC. Pan et al. (34) showed in their study that *HOOK1* was associated with PC and decreased during cancer progression. *EME2* is one of the genes which was identified in this study. We do not find any study that clearly stated that this gene is associated with PC, but there is some evidence of an association between *EME2* and PC on the Human Protein Atlas website (https://www.proteinatlas.org/).

S100P, *GPX3*, *CDH4*, *ITGA2B*, *DTX3*, *ZMIZ2*, *DDX3X*, and *KRAS* were upregulated. As it seems, the expression of the *ACRV1*, *LILRA2*, *CDC14B*, *ENG*, and *LRRK1* is higher in the pancreas group than in the healthy control group, but this difference was not as obvious as in the other upregulated genes. *FTH1* and *CD7* were down-regulated. *EME2*, *PERP*, *APOH*, *GALNT6*, and *SULF2* were upregulated. Tarhan et al. (33) and Alhasan et al. (32) report the upregulation of *GALNT6* and *SULF2*, respectively, in PC. In the study by Kuwae et al. (30) as well as in our study, *APOH* was reported as an upregulated gene. Dasgupta et al. (25) demonstrated that *PERP* had

significantly higher expression in PC cells. There were some limitations in this study. Both the up-regulated and the down-regulated genes are important in the microarray dataset, however, in this study, we only evaluate up-regulated genes. Because we did not have access to an appropriate dataset, we had to use the GEO dataset. We also suggest that researchers evaluate different methods for gene selection and compare the performance of the different methods in selecting effective genes in disease.

Conclusion

This study identified some associated genes with PC that are not detected by conventional methods. Since the associated genes presented in this study were confidential, we suggest the researchers to use the random forest algorithm for the selection process as well as the KNN for imputation if there are MVs in their dataset.

Acknowledgments

We would like to thank the Vice-Chancellor of Research and Technology, Hamadan University of Medical Sciences for approving and supporting the study (No 140006024508). There is no conflict of interest in this study.

Authors' Contributions

N.R., A.R.S., M.F., F.B.; Preparation, Conceptualization, Validation, Investigation, Resources, Writing, Review and Editing, Visualization, Supervision, Project administration, Funding acquisition, and Expanding the manuscript. A.R.S., N.R.; Contributed to the methodology, software, and formal analysis. A.R.S., N.R., M.F.; Writing the drafting of the manuscript. F.B.; Contributed to the data curation process of the manuscript. All authors read and approved the final manuscript.

Reference

1. Luo W, Tao J, Zheng L, Zhang T. Current epidemiology of pancreatic cancer: challenges and opportunities. Chin J Cancer Res. 2020; 32(6): 705-719.
2. Huang J, Lok V, Ngai CH, Zhang L, Yuan J, Lao XQ, et al. Worldwide burden of, risk factors for, and trends in pancreatic cancer. Gastroenterology. 2021; 160(3): 744-754.
3. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin. 2019; 69(1): 7-34.
4. Siri FH, Salehiniya H. Pancreatic cancer in Iran: an epidemiological review. J Gastrointest Cancer. 2020; 51(2): 418-424.
5. Jamali A, Kamgar M, Massarrat S, Sotoudeh M, Larijani B, Adler G, et al. Pancreatic cancer: state of the art and current situation in the Islamic Republic of Iran. Govaresh. 2009; 14(3): 189-197.
6. Senapti R, Shaw K, Mishra S, Mishra D. A novel approach for missing value imputation and classification of microarray dataset. Procedia Eng. 2012; 38: 1067-1071.
7. Breiman L. Random forests. Machine Learning. 2001; 45(1): 5-32.
8. Chiu CC, Chan SY, Wang CC, Wu WS. Missing value imputation for microarray data: a comprehensive comparison study and a web tool. BMC Syst Biol. 2013; 7 Suppl 6: S12.
9. Oh S, Kang DD, Brock GN, Tseng GC. Biological impact of missing-value imputation on downstream analyses of gene expression profiles. Bioinformatics. 2011; 27(1): 78-86.
10. Luengo J, García S, Herrera F. On the choice of the best imputation methods for missing values considering three groups of classification methods. Knowl Inf Syst. 2012; 32(1): 77-108.
11. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001; 17(6): 520-525.
12. Zhang L, Farrell JJ, Zhou H, Elashoff D, Akin D, Park NH, et al. Salivary transcriptomic biomarkers for detection of resectable pancreatic cancer. Gastroenterology. 2010; 138(3): 949-957. e1-7.
13. Bennett DA. How can I deal with missing data in my study? Aust N Z J Public Health. 2001; 25(5): 464-469.
14. Dong Y, Peng CY. Principled missing data methods for researchers. Springerplus. 2013; 2(1): 222.
15. Moorthy K, Mohamad M, Deris S. A review on missing value imputation algorithms for microarray gene expression data. Curr Bioinform. 2014; 9(1): 18-22.
16. Lall R, Robinson T. The MIDAS touch: accurate and scalable missing-data imputation with deep learning. Political Analysis. 2020: 1-18.
17. Patterson TA, Lobenhofer EK, Fulmer-Smentek SB, Collins PJ, Chu TM, Bao W, et al. Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. Nat Biotechnol. 2006; 24(9): 1140-1150.
18. Mutch DM, Berger A, Mansourian R, Rytz A, Roberts MA. The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data. BMC Bioinformatics. 2002; 3: 17.
19. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002; 2(3): 18-22.
20. de Brevern AG, Hazout S, Malpertuy A. Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. BMC Bioinformatics. 2004; 5: 114.
21. Liu H, Shi J, Anandan V, Wang HL, Diehl D, Blansfield J, et al. Reevaluation and identification of the best immunohistochemical panel (pVHL, Maspin, S100P, IMP-3) for ductal adenocarcinoma of the pancreas. Arch Pathol Lab Med. 2012; 136(6): 601-609.
22. Bournet B, Pointreau A, Souque A, Oumouhou N, Muscari F, Lepage B, et al. Gene expression signature of advanced pancreatic ductal adenocarcinoma using low density array on endoscopic ultrasound-guided fine needle aspiration samples. Pancreatology. 2012; 12(1): 27-34.
23. Bardeesy N, DePinho RA. Pancreatic cancer biology and genetics. Nat Rev Cancer. 2002; 2(12): 897-909.
24. Kamisawa T, Wood LD, Itoi T, Takaori K. Pancreatic cancer. Lancet. 2016; 388(10039): 73-85.
25. Dasgupta A, Arneson-Wissink PC, Schmitt RE, Cho DS, Ducharme AM, Hogenson TL, et al. Anticachectic regulator analysis reveals Perp-dependent antitumorigenic properties of 3-methyladenine in pancreatic cancer. JCI Insight. 2022; 7(2): e153842.
26. Park JM, Mau CZ, Chen YC, Su YH, Chen HA, Huang SY, et al. A case-control study in Taiwanese cohort and meta-analysis of serum ferritin in pancreatic cancer. Sci Rep. 2021; 11(1): 21242.
27. Fujiwara K, Ohuchida K, Ohtsuka T, Mizumoto K, Shindo K, Ikenaga N, et al. Migratory activity of CD105+ pancreatic cancer cells is strongly enhanced by pancreatic stellate cells. Pancreas. 2013; 42(8): 1283-1290.
28. Ouyang Y, Pan J, Tai Q, Ju J, Wang H. Transcriptomic changes associated with DKK4 overexpression in pancreatic cancer cells detected by RNA-Seq. Tumour Biol. 2016; 37(8): 10827-38.
29. Li L, Zhang JW, Jenkins G, Xie F, Carlson EE, Fridley BL, et al. Genetic variations associated with gemcitabine treatment outcome in pancreatic cancer. Pharmacogenet Genomics. 2016; 26(12): 527-537.
30. Kuwae Y, Kakehashi A, Wakasa K, Wei M, Yamano S, Ishii N, et al. Paraneoplastic ma antigen-like 1 as a potential prognostic biomarker in human pancreatic ductal adenocarcinoma. Pancreas. 2015; 44(1): 106-115.
31. Hao L, Zhang Q, Qiao HY, Zhao FY, Jiang JY, Huyan LY, et al. TRIM29 alters bioenergetics of pancreatic cancer cells via cooperation of miR-2355-3p and DDX3X recruitment to AK4 transcript. Mol Ther Nucleic Acids. 2021; 24: 579-590.
32. Alhasan SF, Haugk B, Ogle LF, Beale GS, Long A, Burt AD, et al. Sulfatase-2: a prognostic biomarker and candidate therapeutic target in patients with pancreatic ductal adenocarcinoma. Br J Cancer. 2016; 115(7): 797-804.
33. Tarhan YE, Kato T, Jang M, Haga Y, Ueda K, Nakamura Y, et al. Morphological changes, cadherin switching, and growth suppression in pancreatic cancer by GALNT6 knockdown. Neoplasia. 2016; 18(5): 265-272.
34. Pan Z, Li L, Fang Q, Zhang Y, Hu X, Qian Y, et al. Analysis of dynamic molecular networks for pancreatic ductal adenocarcinoma progression. Cancer Cell Int. 2018; 18: 2.