Iranian Society of Animal Science

Journal of Livestock Science and Technologies

Shahid Bahonar University of Kerman

ISSN: 2322-3553 (Print) ISSN: 2322-374X (Online)

Paper type: Original Research

Prediction of ovarian cancer in Holstein cattle using machine learning and microarray data

Mostafa Ghaedrahmati, Farhad Ghafouri-Kesbi, Ahmad Ahmadi*

Department of Animal Science, Faculty of Agriculture, Bu-Ali Sina University, Hamedan, Iran

*Corresponding author,
E-mail address:
f.ghafouri@basu.ac.ir

Received: 16 Apr 2025,
Received in revised form: 15 Jul 2025,
Accepted: 24 Jul 2025,
Published online: 25 Jul 2025,

ORCID
Mostafa Ghaedrahmati
0000-0003-1767-7475
Farhad Ghafouri-Kesbi
0000-0002-2219-055X
Ahmad Ahmadi
0000-0003-0276-9027

Abstract The aim was to the network visualization of genes involved in ovarian cancer in Holstein cattle and assess the performance of machine learning (ML) methods for predicting ovarian cancer using gene expression microarray data. Gene expression data with accession number GSE225981 for healthy and cancer ovarian stromal cells in Holstein cows were obtained from the GEO database. Differentially expressed genes (up and down-regulated genes, DEGs) were identified with online web tool GEO2R. After identifying DEGs and genes associated with ovarian cancer, the Cytoscape software was used to visualize the gene network. Decision tree (DT), Random Forest (RF) and Support Vector Machine (SVM) were used to predict the phenotype (healthy or cancer) from the microarray data. The variable importance feature of RF applying the Gini index was used to select and rank the most important genes in the network. Selected genes were then evaluated to determine their contribution in cancer-related pathways. There were 603 differentially expressed genes (DEGs) of which 327 were up-regulated and 276 were down-regulated. Except for the scenario of 2 samples in training data and 4 samples in test data in which the accuracy of DT was 75%, in other scenarios, the ML methods predicted the phenotypes (healthy or cancer) with the accuracy of 100%. The genes *GPR65*, *RHBDF2*, *TBC1D30*, *DSG2*, *H2AC17*, *AFF3*, *AGMO*, *AURKA*, *CA3* and *CA9* were selected by RF as promising potential markers for diagnosis and prediction of ovarian cancer. A literature survey showed the involvement of these genes in the process and cancerous pathways. In conclusion, the studied ML methods were recommended for analyzing microarray data as showed significant performance in predicting ovarian cancer in Holstein cattle. Also, the variable importance feature of RF can be part of any study on microarray data for identifying important genes, those which are highly correlated with the disease in question.

Keywords: machine learning, microarray, gene, cancer, Gini index

Introduction

Genome damage, which leads to aberrant gene expression, is the underlying cause of most cancers. The major genetic damage that occurs in tumors falls into two categories: 1) chromosomal alterations, such as changes in the number or structure of chromosomes, and 2) changes in the DNA sequence, both of which lead to genetic instability in the cell. Often when DNA is damaged, the body can repair it; unfortunately, in cancer cells, damaged DNA is not repaired. Individuals can also inherit damaged DNA from their parents, and thus inherit a predisposition to cancer (Rashidi et al., 2023).

DOI: 10.22103/jlst.2025.25281.1626

Ovarian cancer is the most lethal gynecological malignancy and the 8th leading cause of cancer death in women around the world (Reid, 2018). Ovarian tumors fall generally into three broad categories: surface epithelial tumors, sex cord-stromal tumors, and germ cell tumors (Yener et al., 2004). In humans, more than 95% of ovarian cancers originate in the epithelial cells on the surface of the ovary (Parrot et al., 2000). Therefore, sex cord-stromal tumors and germ cell tumors would comprise the remainder 5%. Ovarian tumors are common in domestic animals but they are not frequent in cows (DesCôteaux et al., 1989; Švara et al., 2009). In 20913 routine transrectal palpations, their incidence was less than 0.5%. In another study of 302 bovine tumors, 7% affected the genital tract, including ovaries in 4.3% of the cases (DesCôteaux et al., 1989). The cow ovaries have been used as a model system to investigate normal ovarian surface epithelium functions in human (Parrot et al., 2000).

Microarray is one of the outputs of recent developments in DNA technology. Microarray provides a basis to genotype thousands of different loci at a time, which is useful for association and linkage studies to isolate those chromosomal regions that are related to a particular disease (Govindarajan et al., 2012). In addition, it allows to measure differential expression of thousands of genes in different cell types such as healthy and cancer cells. This information has the potential to be analyzed with machine learning (ML) methods to predict and diagnose of cancer. Machine learning is a branch of artificial intelligence that aims to achieve machines that are capable of extracting knowledge (learning) from the environment. According to the definition of machine learning, it is how to write a program that learns through experience and corrects and improves its performance at each stage. A machine learns whenever it can make changes to its structure, program, or information, and therefore, it is expected to make positive changes in its future performance (Nilsson, 1998). Machine learning is used in various topics that involve classification (the machine learns to assign inputs to predetermined categories), clustering (the machine learns to discover which inputs fit together in a category), and prediction (the machine predicts the numerical value of input) (Bishop, 2006). Gene expression data of DNA microarray can be analyzed with ML to either determine whether the patient is oncological or not (two-class problems), distinguish between different types of cancer (multi-class problems), predict the response to a drug based on the gene signature, or identify tumors by finding groups of similarly expressed genes. Compared with traditional methods, applying ML improves the accuracy of cancer prediction by about 60% (Abd-Elnaby et al., 2022). In addition, by using ML, it is possible to identify the most important genes associated with cancer. This is done in a process so-called ‘gene selection’. Gene selection is the technique applied to the gene expression dataset, such as DNA microarray, to reduce the number of genes that are redundant and less expressive or less informative and, therefore, identify the relevant genes for subsequent research (Mahendran et al., 2020). This involves obtaining a set of genes that are related to the outcome of interest. Selected genes could be used as gene markers for the prediction and diagnosis of cancer (Ram et al., 2017).

Recently, we have witnessed lots of activities in application of ML to predict and detect diseases in livestock. For example, Magana et al. (2023) used machine learning algorithms based on sensor behavior data for prediction of dermatitis in dairy cows. The ML which was based on the Tree-Based Pipeline Optimization Tool (TPOT), predicted dermatitis 2 days prior to the appearance of the first clinical signs with an accuracy of 64%. Lasser et al. (2021) used different ML methods to predict a variety of diseases in dairy cows. For anestrus, the accuracy of Logestick Regression, Random Forest and Boosting reached 0.97, 0.97 and 0.95. However, so far, ML has not been applied to predict cancer in livestock by using gene expression microarray data. Therefore, the aims of the present study were 1) visualization of the gene network involved in ovarian cancer in Holstein cattle, 2) to predict ovarian cancer in Holstein cattle from microarray data by decision tree (RT), Random Forest (RF) and Support Vector Machine (SVM) and 3) applying Random Forest to identify most important genes associated with ovarian cancer in Holstein cattle.

Materials and methods

Data

Gene expression data from healthy and tumor ovarian stromal cells (accession number GSE225981) were extracted from the GEO Expression Omnibus database (Clough et al., 2024). The data were generated using the Bovine Gene 1.0 ST Array, which is used to measure the gene expression of 24,415 probes (genes) and included mRNA transcriptome data of normal ovarian stromal cells and tumor ovarian stromal cells. Data were classified into two groups. The first group was animals with ovarian cancer (3 samples) and the second group was healthy animals (control treatment, 3 samples) (Table 1).

Table 1. Accession number of ovarian tissues

GSM7061229	Normal ovary
GSM7061230	Normal ovary
GSM7061231	Normal ovary
GSM7061232	Ovarian cancer
GSM7061233	Ovarian cancer
GSM7061234	Ovarian cancer

Data analysis

Microarray data pre-processing

Before gene expression analysis, quality control of raw data was performed. The Limma package in R (Ritchie et al., 2015) was employed to preprocess data, including background correction, between and within normalization, and final probe summarization. Because of batch effects, the outlier samples should be removed. After the analyses of the samples, no outlier was observed. We utilized GEO2R (http://www.ncbi.nlm.nih.gov/geo/geo2r/) to identify differentially expressed genes (up and down-regulated genes, DEGs). The criteria for identifying the DEGs using GEO2R were (|log2 fold-change (FC)| > 1.5) and adj *P*-value < 0.05. The value of log2 FC is the difference between log2 of the averaged gene expression value in the cancer cells and the log2 of the averaged expression value of that gene in the normal cells. The GEO2R is a web-based tool designed to facilitate the comparison of two or more groups of samples within a GEO dataset, helping researchers pinpoint genes that exhibit differential expression under varying experimental conditions. The GEO2R accomplishes this by examining the raw microarray data provided by the original submitters, employing the GEOquery and limma R packages from the Bioconductor project (Roudbari et al., 2023). After identifying the DEGs and genes associated with ovarian cancer, the STRING database (Jensen et al., 2009) was utilized to retrieve interacting genes/proteins. This database is a reliable tool for obtaining protein interaction network information. For the visualization network, the Cytoscape software (Shannon et al., 2003) was employed, where nodes represent proteins and edges represent their interactions.

<u>Machine learning methods</u>

Decision tree (DT): This algorithm is normally represented in a tree structure. A decision tree is described as a classifier using a recursive split of the instance space. It generates a predictive model that connects node observations to inferences about the desired value of the nodes. The leaf in a tree structure represents the class (Tarawneh et al., 2022). Let $\boldsymbol{y}$ (n×1) be the vector observations, and $\boldsymbol{X} = \{\boldsymbol{x}_i\}$, where $\boldsymbol{x}_i$ is a (p×1) vector representing the expression scores of each animal for p genes. The DT model can be represented as follows:

$$\Psi(\boldsymbol{y}.\boldsymbol{X})$$

1) The DT is constructed as follows: 1) different samples from the training data set, i.e., {(x1, y1), . . . , (xn, yn)}, are drawn with replacement, 2) a small group of the genes is randomly selected from the *p* genes marker and the gene *j* which minimizes the lost function is selected, 3) according to the expression score of the gene *j*, the node is split in two child nodes and individuals go to one of the child nodes, 4) steps 2-3 are repeated until all the terminal nodes become maximally homogeneous. The package *tree* (Riplay, 2024) was used to run DT in R (R Development Core Team, 2024).

Random Forest (RF): Random Forest uses an ensemble of decision trees, grown on bootstrap samples of observations using a random subset of predictors to define the best split at each node. Each node of the tree has access to only one randomly chosen subset of features while training a decision tree in the RF approach. The RF model was as follows:

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x.\Psi_b)$$

The RF prediction for a new observation $x(\hat{f}_{rf}^{B}(x))$, is computed by averaging the predictions over B trees, $\{T(x.\Psi_b)\}_1^B$, for which the given observation was not used to build the tree. Where Ψ_b characterizes the b_{th} RF tree in terms of split variables, cut points at each node, and terminal node values.

Ranking predictor variables with respect to their ability to predict the response is one of the features of RF. The latter was done by considering the so-called “variable importance” measures (VIMs). The Gini index was used as a criterion for computing the importance of each gene in the network and a sub-class including the 10 most important genes was extracted. The package *randomForest* (Liaw and Wiener, 2024) was used to run RF.

Support Vector Machines (SVM): The input are gene expression scores and phenotypic information of animals in the training data $(\boldsymbol{x_i}\,.\,y_i)$ and learn for it a corresponding weight w_i. The output are prediction of unlabeled inputs, i.e., those not in the training set (i.e., the class label of samples ($\hat{y}$)). In SVM, with the input dataset $G = \{(\boldsymbol{x_i}.\,d_i)\}_i^n$ (where $\boldsymbol{x}_i$ is the input vector, d_i is the desired real-valued labeling, and n is the number of input records), $\boldsymbol{x}$ is first mapped into a higher-dimension feature space F via a nonlinear mapping Θ, then linear regression is performed in this space. In other words, SVM approximates a function using the following equation (Hastie et al., 2009):

$$y = f(x) = w\Theta(x) + b$$

The coefficients w and b are estimated by minimizing:

$$R(C) = \frac{1}{2}\|w\|^2 + C\frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(d_i, y_i) \qquad *$$

where L_ε (d, y) is the empirical error measured by ε-insensitive loss function

$$L_\varepsilon(d,y) = \begin{cases} |d-y| - \varepsilon, if\ |d-y| \geq 0 \\ 0, othervise \end{cases}$$

and the term 1/2||w||^2 is a regularization term. The constant C is specified by the user, and it determines the trade-off between the empirical risk and the regularization term. The ε is also specified by the user, and it is equivalent to the approximation accuracy of the training data. The estimates of w and b are obtained by transforming Eq. (*) into the primal function:

$$R(w, \epsilon^{(*)}) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\epsilon_i + \epsilon_i^*)$$

By introducing Lagrange multipliers, the optimization problem can be transformed into a quadratic programming problem. The solution takes the following form:

$$y = f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(x, x_i) + b$$

where K is the kernel function $K(x, xi) = \Theta(x)^T \Theta(xi)$. By using a kernel function, we can deal with problems of arbitrary dimensionality without having to compute the mapping Θ explicitly. Different kernel functions can be selected to map (or transform) input data to feature space. According to Kasnavi *et al.* (2018), we used radial kernels to construct SVM. The package *e1071* (Meyer et al., 2024) was used for SVM analysis.

Predictive performance of ML methods

Accuracy is an important and widely used parameter to evaluate the performance of the models. Accuracy was calculated as the ratio between the correctly predicted instances to the total number of predicted instances.

Results

Figure 1 is the volcano plot. A volcano plot is a helpful visualization that shows the log fold change versus the negative log p-value. It shows up-and down-regulated genes. Usually, we will expect that genes with larger absolute fold changes will have larger negative log p-values hence implying greater statistical confidence. In Figure 1, red points show up-regulated genes and blue points show down-regulated genes. There were 603 differentially expressed genes (DEGs) of which 327 genes were up-regulated and 276 genes were down-regulated. These genes form a network in which they interact with each other (Figure 2).

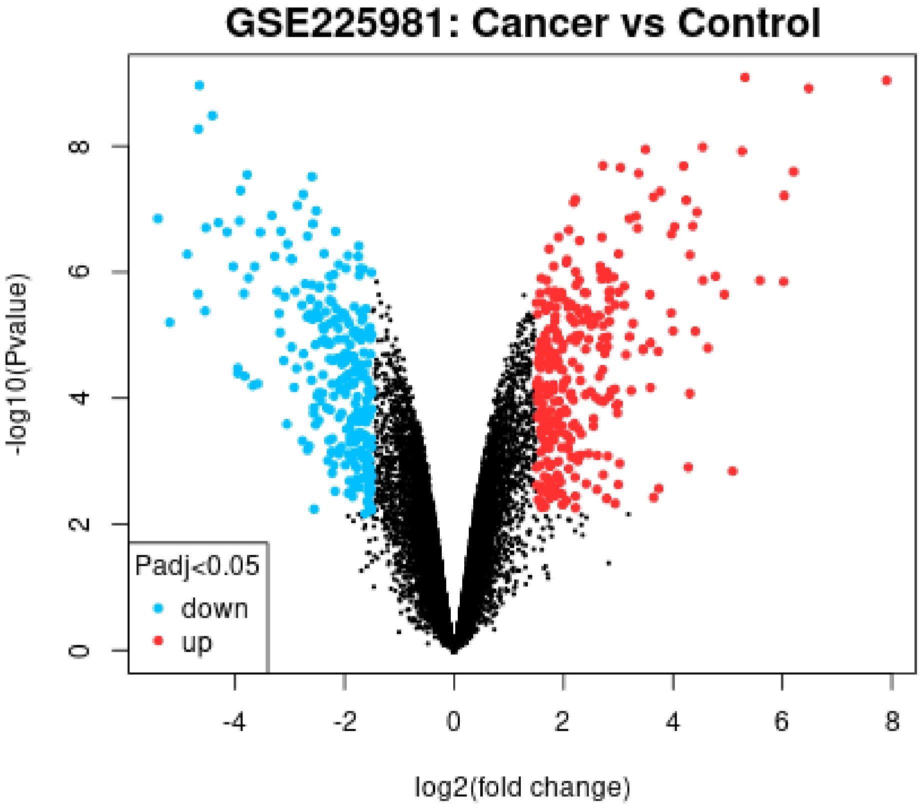

Figure 1. Volcano plot (red dots are up-regulated genes and blue dots ae down-regulated genes)

The results of ML outputs in different scenarios of the number of samples in training and test data are presented in Table 2. Except for the scenario of 2 samples in training data and 4 samples in test data in which the accuracy of DT was 75%, in other scenarios, ML methods predicted the phenotypes (normal or cancer) with the accuracy of 100%.

Figure 3 shows the genes selected by RF according to their importance for predicting phenotype measured by a decrease in the Gini index. As shown *GPR65*, *RHBDF2*, *TBC1D30*, *DSG2*, *H2AC17*, *AFF3*, *AGMO*, *AURKA*, *CA3* and *CA9* genes were ranked as first to tenth.

Table 2. Accuracy of ML methods in prediction of ovarian cancer[a]

Number of samples in training data	Number of samples in test data	DT	RF	SVM
2	4	75%	100%	100%
3	3	100%	100%	100%
4	2	100%	100%	100%
5	1	100%	100%	100%

[a] DT: Decision tree, RF: Random Forest, SVM: Support Vector Machine

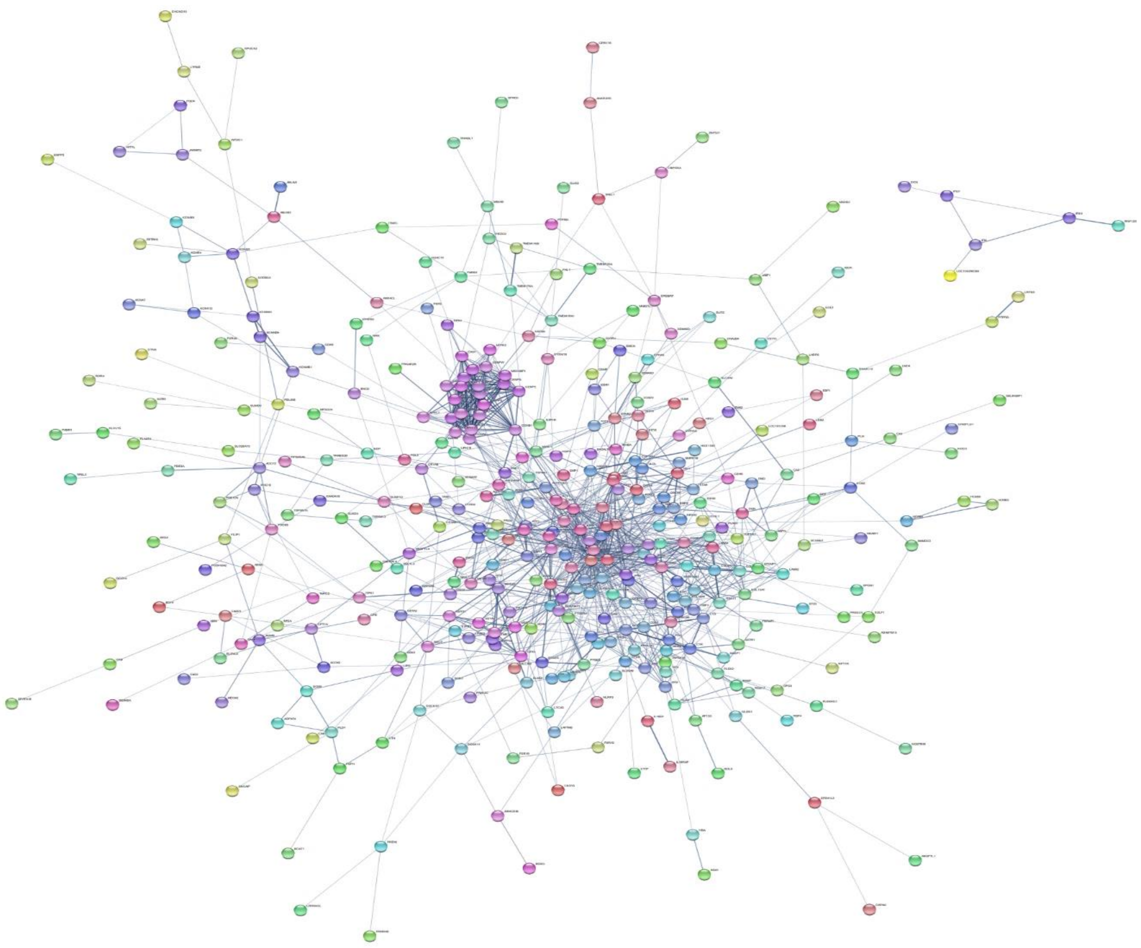

Figure 2. Gene network associated with ovarian cancer in Holstein cow

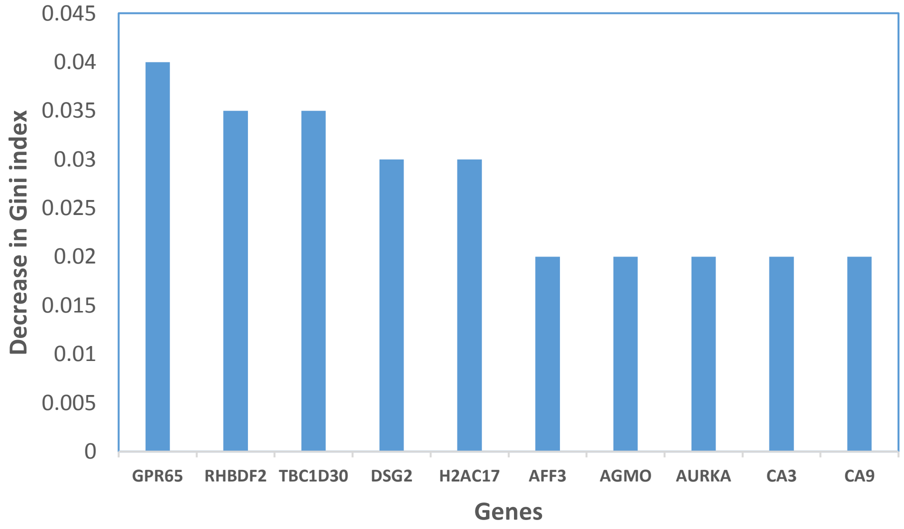

Figure 3. Most important genes associated with ovarian cancer in Holstein cattle selected by Random Forest

Discussion

Cancer is detected using traditional methods, e.g., physical detection, blood test, ultrasound scan, and X-ray scan. For ovarian cancer diagnosis, a blood test and a scan are usually done first, but other tests such as a transvaginal scan are often needed. But they are time-consuming and subject to human errors. Therefore, an effective tool for the diagnosis of breast cancer is necessary, and for this purpose, microarray technology is a promising tool (Abd-Elnaby et al., 2021). Microarray gene expression data has been used to aid in cancer's effective and early detection (Govindarajan et al., 2012; Abd-Elnaby et al., 2021; Gupta et al., 2022; Rezaee et al., 2023). ElAraby et al. (2024) analyzed the data used in this study (GSE225981) to discover the hub genes associated with the prognosis of ovarian cancers. They reported ESR1 and ITGA2 as the most up-regulated and down-regulated genes. In addition, complement and coagulation cascades were the most implicated pathways in ovarian tumor. Here, we investigated the use of DT, RF, and SVM for the classification of microarray data including two classes of healthy and cancer samples. Our results showed excellent performance of ML methods in predicting ovarian cancer from microarray data. The first application of ML in cancer detection and diagnosis can be traced back to the mid-1980s (Simes, 1985). According to the latest PubMed statistics, more than 3500 papers have been published on the subject of using machine learning methods to identify, classify, detect, or distinguish tumors and other malignancies. In most papers we examined, the accuracy of ML methods in predicting cancer from microarray data was high. Rupapara et al. (2022) investigated the possibility of using DNA microarray data for early diagnosis of leukemia. In their study, a logistic regression decision tree (LVtree) was able to distinguish cancer samples from healthy samples with 100% accuracy by reading microarray data. Rezaee et al. (2022) analyzed microarray data related to lymphoma, leukemia, and prostate cancer using k-nearest neighbor (KNN) and deep neural network (DNN). The KNN method was used to select the most important genes in the network, and the selected genes were introduced as input to the deep neural network for cancer type detection. The neural network classified the samples into three groups: lymphoma, leukemia, and prostate cancer with 97%, 99%, and 96% accuracy, respectively. Nogueira et al. (2023) investigated the SVM and DT for predicting and classifying different types of cancer using DNA microarray data. The classification error rate in the SVM method ranged from 00% (ovarian cancer) to 30% (breast cancer). In the DT, the classification error rate ranged from 3% (ovarian cancer) to 40% (breast cancer). In general, the classification error rate increased in cases where the number of samples was low. Alabdulqader et al. (2023) applied a novel weighted convolutional neural network (CNN) model on a 22,283-gene leukemia microarray gene data to predict leukemia cases and reported that the CNN predicted the leukemia cases with a remarkable 99.9 % accuracy. Gupta and Gupta (2021) compared the accuracy of artificial neural networks, Restricted Boltzmann Machine, Deep Autoencoders, and Convolutional Neural Networks (CNN) for post-operative survival analysis of breast cancer patients. The accuracy score achieved by Restricted Boltzmann Machine performed was the highest (0.97), followed by deep Autoencoders that attained an accuracy score of 0.96. CNN achieved a 92% accuracy score, while artificial neural networks attained the least accuracy score (0.89). Nagra et al. (2024) used a new variant of Particle swarm optimization (PSO) method called Self-inertia weight adaptive PSO for microarray cancer classification. The accuracy of classification ranged from 90% (Brain cancer) to 100% (lymphoma). Although previous reports (Gupta and Gupta, 2021; Rupapara et al., 2022; Rezaee et al., 2022; Nogueira et al., 2023; Alabdulqader et al., 2023; Nagra et al., 2024) showed that ML methods can predict different cancer types with high accuracy (up to 100%), it should be noted that in our study the sample size was small (six samples), therefore, further works are needed to validate current findings.

Selection of relevant genes for sample classification is a common task in most gene expression studies, where researchers try to identify the smallest possible set of genes that can still achieve good predictive performance. Various gene selection methods have been developed in the context of machine learning so far. Here we used RF for gene selection because RF has shown excellent performance even when there is noise in the data and can be used when the number of variables is much larger than the number of observations and in problems involving more than two classes. It also returns measures of variable importance (Diaz Uriarte et al., 2006). Jiang et al. (2004) analyzed two microarray gene expression data sets with RF to select lung adenocarcinoma marker genes. They showed excellent performance of variable selection using the RF for their data sets. Among marker genes selected by RF, 7 were found to be cancer-related. Furthermore, based on these marker genes, the RF which was built from one data set predicted the other data set with more than 98% accuracy. In our study, we proposed the 10 most important genes as ovarian cancer biomarkers in Holstein cows. Our literature survey showed that these genes were involved in different cancer types. The GPR65 gene, also known as TDAG8, encodes a proton-sensing G protein-coupled receptor. It is involved in various cellular processes, including immune responses and tumor development. The GPR65 gene is activated by extracellular protons and can modulate downstream signaling pathways. Wang et al. (2023) studied GPR65 genes and found that this gene is differentially expressed in various cancers and linked to tumor mutational burden (TMB), microsatellite instability (MSI), and Ploidy, playing a key function in the tumor microenvironment (TME). They stated that GPR65 could be a target for tumor immunotherapy. The RHBDF2 gene, also known as iRhom2, encodes a protein involved in the regulation of protein secretion, particularly of growth factors and tumor necrosis factor alpha (TNF-α). It is implicated in several human diseases, including familial esophageal cancer and other cancers, and plays a role in processes like epithelial regeneration and wound healing. Saarinen et al. (2012) studied mutations in the RHBDF2 gene and confirmed mutations in the RHBDF2 gene as the underlying cause of the Tylosis with esophageal cancer (TOC) syndrome. The TBC1D3 family is overexpressed in many cancers, including kidney renal clear cell carcinoma (KIRC), which is associated with tumor-infiltrating lymphocytes (Wang et al., 2021). Desmoglein-2 (DSG2 gene) is a calcium-binding single-pass transmembrane glycoprotein and a member of the large cadherin family, which are crucial for maintaining tissue integrity. DSG2 is particularly important in cardiac muscle, where it is found in intercalated discs, and mutations in DSG2 have been linked to heart muscle diseases like arrhythmogenic right ventricular cardiomyopathy (ARVC) and dilated cardiomyopathy (DCM) (Awad et al., 2006). Bioinformatic analyses revealed that DSG2 was significantly up-regulated in cervical cancer compared to normal cervical tissues at both mRNA and protein levels. Up-regulated DSG2 promotes tumor growth and reduces immune infiltration in cervical cancer (Zhang et al., 2024). H2AC17 (H2A Clustered Histone 17) is a protein-coding gene. Diseases associated with H2AC17 include hatologic cancer and plasma cell neoplasm. Among its related pathways are HCMV infection and infectious disease (https://www.genecards.org/). The AFF3, also known as ALF transcription elongation factor 3, plays crucial roles in lymphoid cell development, transcription elongation, protein binding, and various cellular processes. It is a member of a gene family with four paralogs and is known to regulate gene expression related to mesoderm and ectoderm development, as well as mesenchymal cell proliferation, cell adhesion, angiogenesis, cartilage and lens development, and immunoglobulin class switch recombination. AFF3 expression was downregulated in cervical cancer, and its levels were correlated with lymph node metastasis (LNM) (Zhang et al., 2024). Zeng et al. (2022) found significant downregulation of AFF3 in gastric cancer tissues as compared with normal tissues. Aurora kinase A (AURKA) belongs to the family of serine/threonine kinases, whose activation is necessary for cell division processes via regulation of mitosis. It plays a key role in regulating spindle assembly, centrosome duplication, and chromosome segregation, processes essential for accurate cell division (Nikonova et al., 2014). Compared with normal tissues, most tumor types show significantly higher expression of AURKA, except for pancreatic adenocarcinoma, PCPG, skin cutaneous melanoma, and thymoma. AURKB has the lowest expression in kidney chromophobe carcinoma and the highest expression in diffuse large B-cell lymphoma (Du et al., 2021). CA9 is a member of carbonic anhydrases (CAs) family. They are a large family of zinc metalloenzymes that catalyze the reversible hydration of carbon dioxide. They show extensive diversity in tissue distribution and in their subcellular localization. CA9 is induced strongly by hypoxia in several tumor cell lines (https://www.genecards.org/). Turner et al. (2002) showed that expression of CA9 was greater in superficial than invasive bladder tumors.

Conclusion

We identified 603 differentially expressed genes in cancer ovarian stroma cells compared to normal cells. Of 603 DEGs, 327 genes were up-regulated and 276 genes were down-regulated. Machin learning methods showed a strong performance in predicting ovarian cancer using microarray data. The RF and SVM were superior to DT in cases where the number of samples in the training data was low. Most of the selected genes by RF were involved in different types of cancers. These genes can be used as potential markers for diagnosis and prediction of ovarian cancer.

Acknowledgement

This study was supported by Bu-Ali Sina University as a PhD thesis program funding.

Conflict of interest

The authors declare no conflict of interest between authors and other people, institutions and organizations.

References

Abdelwahab, O., Awad, N., Elserafy, M., Badr, E.A., 2022. Feature selection-based framework to identify biomarkers for cancer diagnosis: A focus on lung adenocarcinoma. *PLoS ONE* 17, e0269126.

Alabdulqader, E.A., Alarfaj, A.A., Umer, M., Eshmawi, A., Alsubai, S., Kim, T., Ashraf, I., 2024. Improving prediction of blood cancer using leukemia microarray gene data and chi2 features with weighted convolutional neural network. *Scientific Reports* 14, 15625.

Awad, M.A., Dalal, D., Cho, E., 2006. *DSG2* mutations contribute to arrhythmogenic right ventricular dysplasia/cardiomyopathy. *The American Journal of Human Genetics* 79, 136-142.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer, New York, USA.

Clough, E., Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Sherman, P.M., 2024. NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update. *Nucleic Acid Research* 52, 138-144.

DesCôteaux, L., Harvey, D., Girard, C., 1989. Tumeur des cellules de la granulosa chez une taure: observations cliniques, endocrinologiques et post-mortem. *Canadian Veterinary Journal* 30, 501-503.

Díaz-Uriarte, R., De Andres, S.A., 2006. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7, 3.

ElAraby, I.E., Saleh, A.A., Zaghlol, A.W., Hussien, N., 2024. Bioinformatics study of the microarray data set of ovarian carcinoma in cattle. *Zagazig Veterinary Journal* 52, 367-379.

Gupta, S., Gupta, M.K., 2021. A comparative analysis of deep learning approaches for predicting breast cancer survivability. *Archives of Computational Methods in Engineering* 29, 2959-2975.

Gupta, S., Gupta, M.K., Shabaz, M., Sharma, A., 2022. Deep learning techniques for cancer classification using microarray gene expression data. *Frontiers in Physiology* 13, 952709.

Hastie, T.J., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. 2nd Ed. Springer, New York, USA.

Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Simonovic, M., 2009. STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acid Research* 37, 412-416.

Jiang, H., Deng, Y., Chen, H.S., Tao, L., Sha, Q., Chen, J., Tsai, C.J., Zhang, S., 2004. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics* 5, 81.

Govindarajan, R., Duraiyan, J., Kaliyappan, K., Palanisamy, M., 2012. Microarray and its applications. *Journal of Pharmacy and Bioallied Sciences* 4, S310-2.

Lasser, J., Matzhold, C., Egger-Danner, C., Fuerst-Waltl, B., Steininger, F., Wittek, T., 2021. Integrating diverse data sources to predict disease risk in dairy cattle-a machine learning approach. *Journal of Animal Science* 99, 1-14.

Liaw, A., Wiener, M., 2024. Breiman and Cutler's random forests for classification and regression. Available at http://cran.r-project.org/web/packages/randomForest/index.html.

Magana, J., Gavojdian, D., Menahem, Y., Lazebnik, T., Zamansky, A., dams-Progar, A., 2023. Machine learning approaches to predict and detect early-onset of digital dermatitis in dairy cows using sensor data. *Frontiers in Veterinary Science* 10, 1295430.

Mahendran, N., Durai Raj Vincent, P.M., Srinivasan, K., Chang, C.Y., 2020. Machine learning based computational gene selection models: A survey, performance evaluation, open issues, and future research directions. *Frontiers in Genetics* 11, 603808.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, K., 2024. Misc functions of the department of statistics (e1071), TU Wien. Available at http://cran.r-project.org/web/packages/e1071/index.html.

Nagra, A.A., Khan, A.H., Abubakar, M., Faheem, M., Rasool, A., Masood, K., Hussain, M., 2024. A gene selection algorithm for microarray cancer classification using an improved particle swarm Optimization. *Scientific Reports* 14, 9613.

Nikonova, A.S., Astsaturov, I., Serebriiskii, I.G., Dunbrack, R.L., Golemis, E.A., 2013. Aurora A kinase (AURKA) in normal and pathological cell division. *Cell and Molecular Life Science* 70, 661-687.

Nilsson, N.J., 1998. Introduction to machine learning. Stanford University. Stanford, USA.

Parrot, J.A., Kim., G., Skinner, K., 2000. Expression and action of kit ligand/stem cell factor in normal human and bovine ovarian surface epithelium and ovarian cancer. *Biology Reports* 62, 1600-1609.

Ram, M., Najafi, A., Shakeri, M.T., 2017. Classification and biomarker genes selection for cancer gene expression data using random forest. *Iranian Journal of Pathology* 12, 339-347.

Rashidi, S., Asadi, A., Abdolmaleki, A., 2021. Cancer stem cells: a narrative review. *Journal of Rafsanjan University of Medical Science* 20, 226.

R Development Core Team., 2024. R: A language and environment for statistical computing. Vienna, Austria. Available at: https://www.R-project.org/.

Reid, B.M., Permuth, J.B., Sellers TA., 2017. Epidemiology of ovarian cancer: a review. *Cancer Biology and Medicine*14, 9-32.

Rezaee, K., Jeon, G., Khosravi, M.R., Attar, H.H., Sabzevari, A., 2022. Deep learning-based microarray cancer classification and ensemble gene selection approach. *IET System Biology* 16, 120-131.

Ripley, B., 2024. Package tree: classification and regression tree. Available at: https://cran.r-project.org/web/packages/tree/index.html.

Ritchie, M.E., Phipson, B., Wu, D., 2015. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43, e47.

Roudbari, Z., Mokhtari, M., Ebrahimpour Gorji, A., Sadkowski, T., Sadr, A.S., Shirali, M., 2014. Identification of hub genes and target miRNAs crucial for milk production in Holstein Friesian dairy cattle. *Genes* 14, 2015.

Rupapara, V., Rustam, A., Amaar, P.B., Washington, E.L., Ashraf, I., 2021. Deepfake tweets classification using stacked bi-LSTM and words embedding. *Peer Journal of Computer Science* 7, e745.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J. T., Ramage, D., Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks .*Genome Research* 13, 2498-2504.

Saarinen, S., Vahteristo, P., Lehtonen, R., Aittomaki, K., Launonen, V., Kiviluoto, T., Aaltonen L.A., 2012. Analysis of a Finnish family confirms RHBDF2 mutations as the underlying factor in Tylosis with esophageal cancer. *Familial Cancer* 11, 525-528.

Svara, T., Gombač, M., Juntes, P., Pogačnik, M., 2009. Malignant ovarian granulosa cell tumour in a ewe. *Acta Veterinaria Brno* 78, 281-285.

Simes, R.J., 1985. Treatment selection for cancer patients: application of statistical decision theory to the treatment of advanced ovarian cancer. *Journal of Chronic Diseases* 38, 171-86.

Tarawneh, O., Otair, M., Husni, M., Abuaddous, H.Y., Tarawneh, M., Almomani, M.A., 2022. Breast cancer classification using decision tree algorithms. *International Journal of Advanced Science and Computer* 13, 676.680.

Turner, K.J., Crew, J.P., Wykoff, C.C., Watson, P.H., Poulsom, R., Pastorek, J., Ratcliffe, P.J., Cranston, D., Harris, A.L., 2002. The hypoxia-inducible genes VEGF and CA9 are differentially regulated in superficial vs invasive bladder cancer. *British Journal of Cancer* 86, 1276-1282.

Wang, B., Chen, D., Hua, H., 2021. TBC1D3 family is a prognostic biomarker and correlates with immune infiltration in kidney renal clear cell carcinoma. *Molecular Oncology* 22, 528-538.

Wang, L., Sun, L., Sun, H., Xing, Y., Zhou, S., An, G., 2023. eGPR65 as a potential immune checkpoint regulates the immune microenvironment according to pan-cancer analysis. *Heliyon* 9, e13617.

Zhang, G., Chen, Z., Wang, Y., Huang, A., Nie, F., Gao, L., 2024. Up-regulated DSG2 promotes tumor growth and reduces immune infiltration in cervical cancer. *Pathology-Research and Practical* 262, 15555.

Zeng, Y., Zhang, X., Li, F., Wang., Y., Wei, M., 2022. AFF3 is a novel prognostic biomarker and a potential target for immunotherapy in gastric cancer. *Journal of Clinical Laboratory Analysis* 36, e24437.