



## OPEN ACCESS

## EDITED BY

Ayan Banerjee,  
Arizona State University, United States

## REVIEWED BY

Saurav Mallik,  
Harvard University, United States  
Imane Lamrani,  
Nikola Corporation, United States  
Payal Kamboj,  
Arizona State University, United States, in  
collaboration with reviewer [IL]

## \*CORRESPONDENCE

Neda Gilani

✉ neda.gilani@gmail.com

Reza Arabi Belaghi

✉ rezaarabi11@gmail.com

RECEIVED 16 March 2023

ACCEPTED 20 July 2023

PUBLISHED 09 August 2023

## CITATION

Hamidi F, Gilani N, Arabi Belaghi R, Yaghoobi H,  
Babaei E, Sarbakhsh P and Malakouti J (2023)  
Identifying potential circulating miRNA  
biomarkers for the diagnosis and prediction of  
ovarian cancer using machine-learning  
approach: application of Boruta.  
Front. Digit. Health 5:1187578.  
doi: 10.3389/fdgth.2023.1187578

## COPYRIGHT

© 2023 Hamidi, Gilani, Arabi Belaghi, Yaghoobi,  
Babaei, Sarbakhsh and Malakouti. This is an  
open-access article distributed under the terms  
of the [Creative Commons Attribution License  
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in  
other forums is permitted, provided the original  
author(s) and the copyright owner(s) are  
credited and that the original publication in this  
journal is cited, in accordance with accepted  
academic practice. No use, distribution or  
reproduction is permitted which does not  
comply with these terms.

# Identifying potential circulating miRNA biomarkers for the diagnosis and prediction of ovarian cancer using machine-learning approach: application of Boruta

Farzaneh Hamidi<sup>1</sup>, Neda Gilani<sup>1,2\*</sup>, Reza Arabi Belaghi<sup>3,4,5\*</sup>,  
Hanif Yaghoobi<sup>6</sup>, Esmaeil Babaei<sup>6,7</sup>, Parvin Sarbakhsh<sup>1</sup>  
and Jamileh Malakouti<sup>8</sup>

<sup>1</sup>Department of Statistics and Epidemiology, Faculty of Health, Tabriz University of Medical Sciences, Tabriz, Iran, <sup>2</sup>Road Traffic Injury Research Center, Tabriz University of Medical Sciences, Tabriz, Iran, <sup>3</sup>Department of Mathematics, Applied Mathematics and Statistics, Uppsala University, Uppsala, Sweden, <sup>4</sup>Department of Statistics, Faculty of Mathematical Science, University of Tabriz, Tabriz, Iran, <sup>5</sup>Department of Energy and Technology, Swedish Agricultural University, Uppsala, Sweden, <sup>6</sup>Department of Biological Sciences, School of Natural Sciences, University of Tabriz, Tabriz, Iran, <sup>7</sup>Interfaculty Institute for Bioinformatics and Medical Informatics (IBMI), University of Tübingen, Tübingen, Germany, <sup>8</sup>Department of Midwifery, Faculty of Nursing and Midwifery, Tabriz University of Medical Science, Tabriz, Iran

**Introduction:** In gynecologic oncology, ovarian cancer is a great clinical challenge. Because of the lack of typical symptoms and effective biomarkers for noninvasive screening, most patients develop advanced-stage ovarian cancer by the time of diagnosis. MicroRNAs (miRNAs) are a type of non-coding RNA molecule that has been linked to human cancers. Specifying diagnostic biomarkers to determine non-cancer and cancer samples is difficult.

**Methods:** By using Boruta, a novel random forest-based feature selection in the machine-learning techniques, we aimed to identify biomarkers associated with ovarian cancer using cancerous and non-cancer samples from the Gene Expression Omnibus (GEO) database: GSE106817. In this study, we used two independent GEO data sets as external validation, including GSE113486 and GSE113740. We utilized five state-of-the-art machine-learning algorithms for classification: logistic regression, random forest, decision trees, artificial neural networks, and XGBoost.

**Results:** Four models discovered in GSE113486 had an AUC of 100%, three in GSE113740 with AUC of over 94%, and four in GSE113486 with AUC of over 94%. We identified 10 miRNAs to distinguish ovarian cancer cases from normal controls: hsa-miR-1290, hsa-miR-1233-5p, hsa-miR-1914-5p, hsa-miR-1469, hsa-miR-4675, hsa-miR-1228-5p, hsa-miR-3184-5p, hsa-miR-6784-5p, hsa-miR-6800-5p, and hsa-miR-5100. Our findings suggest that miRNAs could be used as possible biomarkers for ovarian cancer screening, for possible intervention.

## KEYWORDS

artificial intelligence, Boruta, biomarker, feature selection, Gene Expression Omnibus, ovarian cancer, oncology

## 1. Introduction

Ovarian cancer is most often found in granulosa cells or germ cells, with epithelial histology accounting for more than 90% of all ovarian cancer. Epithelial ovarian cancer (EOC) (1) is a widespread gynecologic malignancy in industrialized and developing countries (2), with approximately 230,000 new cases and nearly 140,000 deaths per year

(3). In 2020, the United States was expected to see 21,750 new cases and 13,940 deaths (4), while Europe experienced 29,000 deaths (5). According to the International Federation of Gynecology and Obstetrics (FIGO), only 30% of advanced-stage cancer patients live for nearly 5 years after receiving a primary-stage prognosis (6, 7). Only 19% of ovarian cancer patients are diagnosed at its early stage due to the absence of robust and minimally invasive methods at its early detection (8). Hence, advanced approaches for the early screening of ovarian cancer are necessary for proper medication and timely treatment. Regarding the genetic basis of cancer malignancy, microarray technology (9) has recently been one of the most widely used tools to evaluate the functions of genes in related patients. MicroRNAs (miRNAs) are short (18–25 nucleotides in length) non-coding RNAs that have emerged as important translational gene regulators in cancer cells (6). The screening models currently available are insufficient, and accurate non-invasive molecular biomarkers are urgently needed. Many studies have looked at the expression profiles of miRNAs in tissue and serum samples from ovarian cancer patients to identify appropriate biomarkers (10). Even though in many studies miRNAs are still insufficient for clinical applications that are due to large-scale non-validation and inconsistencies in the diagnosis of devices (11–13), it could expand a new screening strategy that can differentiate cancerous from non-cancerous women. In addition, the comprehensive characteristics of circulating miRNAs enable us to produce optimal diagnostic models for ovarian cancer (11–14).

## 1.1. Related works

MicroRNA molecules can act as an important tool for the detection of ovarian cancer. Chung et al. (15) reported let-7b, miR-26a, miR-132, and miR-145 as potential biomarkers in ovarian cancer patients. Among the results of Yuan et al.'s (16) study, has-miR-6784-5p, has-miR-6800-5p, and has-miR-5100 are indicating ovarian-associated cancer signature. Jeon et al. (17) reported that the serum and tissue miR-1290 was significantly elevated in patients with epithelial ovarian cancer compared with patients with benign ovarian neoplasm. Chen et al. (18) reported a total of 19 miRNAs, which were identified by random forest models, that were important in cancer diagnosis. In this study, the top five miRNAs with the highest frequency were chosen to be the biomarker candidates for cancer screening, which has-miR-3184-5p achieved a high rank. Yaghoobi et al. (19) proposed a method called EBST that has identified 11 serum miRNAs as potential biomarkers associated with ovarian cancer; among the miRNAs set, has-miR-1228-5p and has-miR-6784-5p were also reported. Zhang et al. (20) reported the four miRNA models that showed very strong performances with AUCs > 0.95 in the biliary tract, bladder, colorectal, esophageal, gastric, glioma, liver, ovarian, pancreatic, and prostate cancers. This study provides proof-of-concept data in demonstrating that the four miRNA (hsa-miR-5100, hsa-miR-1343-3p, hsa-miR-1290, and hsa-miR-4787-3p) model has the potential to be developed into a simple, inexpensive, and non-invasive blood test for the early detection of

multiple cancers with high accuracy. Using statistical approaches, Hamidi et al. (21) identified 10 miRNAs regulated in ovarian serum cancer samples compared with non-cancer samples in the publicly available data set GSE106817: hsa-miR-5100, hsa-miR-6800-5p, hsa-miR-1233-5p, hsa-miR-4532, hsa-miR-4783-3p, hsa-miR-4787-3p, hsa-miR-1228-5p, hsa-miR-1290, hsa-miR-3184-5p, and hsa-miR-320b. However, the approach of the previous study (21) failed to take into account the non-linearity structure in big data; therefore, in this paper, we are implementing a new machine-learning variable selection approach called Boruta to address this problem. We will observe that the new miRNAs will be explored by the new method that has not been recognized in the traditional methods.

## 1.2. Novel contributions

It is important to note that the choice of feature selection (FS) method should be tailored to the specific characteristics of the data set and research question at hand. Gene expression data are the representation of non-linear interactions among genes (22). By computing analysis of these data, it is expected to gain knowledge of gene functions and disease mechanisms. Statistical methods can only identify linear patterns, while non-linear patterns of relationships remain hidden. As mentioned in many research (23–29), Boruta has superior advantages in terms of feature selection accuracy, stability, and classification performance across different domains such as protein subcellular localization and credit risk assessment, however, especially in microarray data sets of ovarian cancer that have been rarely used before. This is based on some studies on the stability of Boruta (30–32) as a machine-learning method that can more accurately discover new miRNAs that were hidden in statistical methods. Therefore, this work attempts an innovation in two important issues: the identification of new miRNAs based on complex non-linear structures and the comparison of new results with the previous ones, which will be described in the results and discussion section.

## 2. Materials and methods

To identify a robust circulating miRNA biomarker, we searched the Gene Expression Omnibus (GEO) database with specific keywords, namely, (“ovarian neoplasms” [MeSH Terms] OR ovarian cancer [All Fields]) AND “Homo sapiens” [porgn] AND “MicroRNAs” [MeSH Terms] OR miRNA [All Fields]. Then, three data sets using the same platform (3D-Gene Human miRNA V21\_1.0.0) with a larger sample size GSE106817, GSE113486, and GSE113740 were included (385 ovarian cancer patients and 3,026 non-cancer controls in total) for further analysis. The GSE106817 has 320 ovarian cancer patients with an average age of 52 years and 2,759 non-cancer controls that were used as the internal discovery data set, and the GSE113486 has 40 ovarian cancer patients and 52 non-cancer controls. The GSE113740 has 25 ovarian cancer patients, and 215 non-cancer

controls were used for independent validation data sets. This study was approved by the ethics committee of Tabriz University of Medical Sciences (no.: IR.TBZMED.REC.1400.006).

## 2.1. Study design and data set

We have used the GSE106817, GSE113486, and GSE113740 data sets from the GEO database, which is available at <https://www.ncbi.nlm.nih.gov/geo/>. The GSE106817 data set started on 13 November 2017 in Kanagawa, Japan, which is serum miRNA profiles of 4,046 women specimens, and which consists of 333 ovarian cancer and 2,759 non-cancer controls and 976 other types of cancer. The GSE106817 data set consists of ovarian cancer patients who were of mean age 57( $\pm$ 12) years, 25% stage I, 10% stage II, 55% serous, 19% clear cell, and 13% endometrioid histology (33). Three microarray data sets totaling to 6,835 unique participants including 728 ovarian cancer patients and 3,892 non-cancer controls were included in the current analysis, all derived from studies originating from a Japanese nationwide research project “Development and Diagnostic Technology for Detection of miRNA in Body Fluids” that is designed to characterize serum miRNAs in over 5,000 participants across several types of cancer using a standardized microarray platform. **Supplementary Figure S1** clearly shows the stages of data pre-processing, identification of significant features or predictors, the model building of classifier algorithms, and performance evaluation, which are the four main phases of this analysis.

### 2.1.1. Participants and serum samples

The serum sample collection has been previously described in the original publications (33–35). Briefly, serum samples were collected from cancer patients who were referred or admitted to the National Cancer Center Hospital (NCCH) and stored at 4°C for 1 week before being stored at –20°C until further use. Cancer patients who were treated with preoperative chemotherapy and radiotherapy before serum collection were excluded. The serum samples for non-cancer controls who had no history of cancer and no hospitalization during the previous 3 months were collected along with routine blood tests from outpatient departments of three sources: NCCH, National Center for Geriatrics and Gerontology (NCGG) Biobank, and Yokohama Minoru Clinic (YMC). Serums collected from NCCH were stored in the same way as the serum from cancer patients, while those from NCGG and YMC were stored at –80°C until use. The original studies were approved by the NCCH Institutional Review Board, the Ethics and Conflict of Interest Committee of the NCGG, and the Research Ethics Committee of Medical Corporation Shintokai YMC. Written informed consent was obtained from each participant.

### 2.1.2. MiRNA microarray expression analysis

The details about microarray analysis were described in the original publications (33–35). Briefly, total RNA was extracted from a 300  $\mu$ l serum, labeled by 3D-Gene<sup>®</sup> miRNA labeling kit

and hybridized to 3D-Gene<sup>®</sup> Human miRNA Oligo Chip (Toray Industries, Kanagawa, Japan) that is designed to investigate 2,588 miRNA sequences registered in miRBase release 21 (<http://www.mirbase.org/>, accessed on 10 January 2022). The following low-quality samples were excluded: coefficient of variation of negative control probes of >0.15 and number of flagged probes identified by 3D-Gene<sup>®</sup> Scanner as “uneven spot images” of >10. The presence of a miRNA was determined when signal intensity was greater than the mean plus two times the standard deviation of the negative control signals, and in using the negative control signals, the top and bottom 5% of the ranked signal intensities were removed. Background subtraction was performed by subtracting the mean signal of negative control signals (after removing the top and bottom 5% as ranked by signal intensities) from the miRNA signal.

## 2.2. Machine learning

In cancer prediction models, statistical and machine-learning algorithms have been widely used, providing more accurate prognoses and lower per-patient costs. The high dimensionality of the gene expression profiles is a crucial issue when building cancer-predictive models (36). As a result, we used a machine-learning algorithm based on the random forest classifier, which is easily implemented in the Boruta package in R (37). In many studies involving miRNAs expression data, Boruta has been used to identify important features (38); this could help in the development of biomarkers for cancer diagnosis and prognosis. On the other hand, we used these techniques to characterize miRNAs with biomarker potential that may be useful in the diagnosis and/or prognosis of this disease, potentially assisting public health (39).

## 2.3. Data cleaning and feature selection

We cleaned and normalized the data using the min-max normalization method (40). Since gene expression data sets had too many irrelevant features for classification, feature selection was inevitable. Feature selection techniques can be used in data pre-processing to perform successful data reduction, which is beneficial for finding accurate data models (41). As noted, feature selection techniques have the benefits of reducing over-fitting and reducing model complexity with ease of understanding, as well as training models more quickly.

### 2.3.1. Boruta

Boruta is a wrapper-based feature selection algorithm that implements a random forest algorithm to iteratively delete the statistically irrelevant features. Boruta searches for all features that are either strongly or weakly relevant to the output variable (27).

Boruta algorithm selects features as follows:

- (a) It assigns randomness to the data set by making shuffled copies of all features (termed as shadow features).

- (b) Next, Boruta uses the data set for training a random forest classifier and uses a feature ranking measure (mean decrease accuracy, MDA) to estimate the relationship with each feature (higher mean value).
- (c) It determines whether a real feature has higher rank than the best of its shadow features on each iteration (in our analysis, 100) and excludes features that are considered extremely insignificant.
- (d) Boruta algorithm comes to a halt when all features have been confirmed.

This would ultimately result in at least a subset of features that is ideal. Since this approach reduces the error of the random forest model, it identifies all features that are either highly significant or unrelated (32, 42, 43). Boruta is used in such a way that the features selected are mostly correlated with the prediction variable.

In the process of identifying if a feature is important or not, some features may be signed by Boruta as “Tentative.” Tentative attributes are decided as confirmed or rejected by using the median Z score of the attributes with the median Z score of the best shadow attribute.

## 2.4. Model building and potential miRNAs signature identification

We split the data using the CARET package into two parts: two-thirds of the data were used for model development or training, while the remaining one-third of the data were used to evaluate or validate the model.

### 2.4.1. Handling of imbalanced classes

In most cases, prediction algorithms train to predict the majority class (i.e., non-cancer), resulting in incorrect sensitivities and specificities (44). Instead, fixing the imbalance in the outcomes (i.e., lower cancer rates) in the training data usually leads to the creation of a better prediction model and a better trade-off between sensitivity and specificity (45). Oversampling the minority class and under-sampling the majority class are the most effective strategy for overcoming imbalanced outcomes (46). To balance the training sample in this article, we used SMOTE random oversampling (47).

### 2.4.2. Find optimal hyperparameters and proposed models

We used a five-fold cross-validation (CV) in the training data set to reduce training errors and obtained the optimal hyperparameters in machine-learning algorithms (48). We performed cancer classification using logistic regression, artificial neural network, decision trees, random forest, and XGBoost (49) algorithms, and to build our models, we applied the varImp() function for finding the most important feature (in our study >80% importance) from each of the proposed models. A brief description of classifiers and their settings are given below or in references therein.

#### 2.4.2.1. Logistic regression

Logistic regression (LR) is used when the answer of a feature is computed as numerical (quantitative) data. The relationship between multiple independent variables and a single binary dependent variable, which is a two-category variable, is investigated using logistic regression. In cancer microarray data, which is a form of the data set in which the outcome (cancer) is determined by the combined outcome of many features (genes), logistic regression has a variety of uses. Logistic regression rejects a linear relationship between the dependent and independent variables in favor of the binomial probability principle, which states that there are only two possible outcomes (50). The fit of a logistic regression model will be evaluated using the area under the curve (AUC) (51).

#### 2.4.2.2. Decision trees

Decision trees (DTs) are a type of supervised machine learning that can be used to find attributes and extract patterns in big databases that are important for predictive modeling (46). The interoperability of the rendered model is a feature of decision tree modeling that distinguishes it from other techniques of pattern recognition. The most straightforward algorithm for processing a visual representation of the relationship between independent and dependent variables is decision trees (52). DTs are easy to build, train, interpret, and explain. However, the variation in the decision trees, in some instances, can be improved using random forests as the outcomes of randomly generated decision trees to produce a more impressive model.

#### 2.4.2.3. Random forest

Random forest (RF) is a supervised ensemble learning algorithm that provides a single combination of prediction accuracy and model interoperability among general machine-learning technique (39). RFs are an instance of ensemble learning, in which a complex model was developed by combining numerous simple decision tree algorithms, due to lower variance than single decision trees. Random forest is a meta-classification approach that fits a number of sub-classifiers (DTs) on various subsets of a data set, and the averages from each decision tree are used to ameliorate the accuracy of classification, the superiorities of RF that they decrease the over-fitting, thus improving accuracy. Random forests can be used to rate the importance of variables in a regression or classification problem (53).

#### 2.4.2.4. Artificial neural networks

In medical research, artificial neural networks (ANNs) have been widely employed (54, 55). When there are complex and non-linear relationships between variables, such algorithms work well. In a word, ANN takes predictors as inputs and connects them to multiple hidden layer combinations with appropriate weights to predict the outcome. The analyst must intelligently choose the hidden layers and weights (56).

#### 2.4.2.5. XGBoosting

Extreme gradient boosting is abbreviated as XGBoost (XGB). XGB is a decision-tree-based ensemble machine-learning algorithm that employs a scalable gradient boosting technique (57). XGB is a

scalable machine-learning system for tree boosting. The most significant component of the success of XGBoost is its scalability across all scenarios. XGB scalability is due to a number of major systems and algorithmic enhancements, parallel and distributed computing speed up learning, allowing for more rapid model exploration. XGB also allows data scientists to process by utilizing out-of-core processing (53).

## 2.5. Evaluation criteria

The validation technique is widely used to avoid over-fitting and to check the validity of the models. We evaluated our outcomes employing two external data sets, as shown in the **Supplementary Figure S1**. The metrics utilized to assess the results of the classification models are expressed below:

$$\text{Accuracy: ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}},$$

$$\text{Sensitivity: SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$\text{Specificity: SPC} = \frac{\text{TN}}{\text{TN} + \text{FP}},$$

$$\text{Kappa: } k = \frac{\text{Pr}(a) - \text{Pr}(e)}{1 - \text{Pr}(e)}$$

where:

1. TP (true positive) is the number of people who suffer from “cancer” among those who were diagnosed with “cancer.”
2. FP (false positive) depicts the number of persons who are “cancerous” but were diagnosed as “non-cancerous.”
3. FN (false negative) is the number of people wrongly found to be “non-cancerous.”
4. TN (true negative) states the number of “non-cancerous” correctly.
5. Pr(a) represents the observed agreement, and Pr(e) represents the chance agreement.

We tested classifier reliability for multi-class data sets using Kappa values, which reflect the compromise among real and expected values (58); positive predictive value (PPV) and negative predictive value (NPV) were also obtained (59). The one-sided DeLong’s test was used to calculate the power for the ROC curves, which was done using the R package “pROC” (60).

## 3. Result

The data have 2,568 variables. In this initial variable section stage by Boruta, 199 variables were selected in 29 min. The training set included 2,156 samples, while the testing set included 923 samples. The training set consisted of 1,932 non-cancerous samples and 224 cancerous samples. After balancing the training data, the non-cancerous and cancerous samples became 1,121 and 1,035, respectively. The data set with reduced

features is classified using LR (statistical), DT and RF (tree-based), ANN, and XGB (machine learning) classifiers. After finding the more important features (in our study over 80%) as shown in **Supplementary Table S1**, we identified 10 potential miRNAs, has-miR-1290, has-miR-1233-5p, has-miR-1914-5p, has-miR-1469, has-miR-4675, has-miR-1228-5p, has-miR-3184-5p, has-miR-6784-5p, has-miR-6800-5p, and has-miR-5100, from the GSE106817 data sets and were defined as the candidate miRNAs for ovarian cancer diagnosis. In **Supplementary Table S2**, we reported the *t*-test table to compare cancer and non-cancerous samples, and all of these miRNAs had significant *P*-value. Using the 10 selected miRNAs, the final machine-learning models with optimal hyperparameters are presented in **Table 1**.

### 3.1. Internal validation data set

As noted in the previous section, we find 10 miRNAs that are has-miR-1290, has-miR-1233-5p, has-miR-1914-5p, has-miR-1469, has-miR-4675, has-miR-1228-5p, has-miR-3184-5p, has-miR-6784-5p, has-miR-6800-5p, and has-miR-5100. We implemented each miRNA separately in models to get their power of prediction individually in classification between cancer and non-cancerous samples. The AUC of each of these miRNAs is listed in **Supplementary Table S1A**. We observe that in the internal validation, all miRNAs have high AUC (minimum AUC: 86.0%; maximum AUC is 96.8%). The performance measures for LR, DT, RF, ANN, and XGB models are shown in **Supplementary Table S3A**. We observe that the AUC of LR, RF, ANN, and XGB is 99.9%. **Supplementary Table S3A** shows the accuracy, sensitivity, specificity, NPV, PPV, and Kappa for LR, DT, RF, ANN, and XGB models in the classification and prediction of ovarian cancer. Four models obtained an AUC of 99.9%; however, DT obtained 98% AUC. In detail, RF has the highest value of accuracy (99.13), specificity (99.51), PPV (95.83), and Kappa (95.35), and LR have high sensitivity (98.96) and NPV (99.88). **Figure 1A** illustrates the ROC curve for the proposed models of 10 candidate miRNAs that are shown in **Supplementary Table S1A**. All models except DT have over 99.9% of AUC. **Figure 1B** shows the individual AUCs of 10 miRNAs in internal data set: has-mir-5100 (93.7%), has-mir-6800-5p (97%), has-mir-6784-5p (94.2%), has-mir-3184-5p (94.2%), has-mir-1228-5p (95.6%), has-mir-4675 (95.4%), has-mir-1469 (96.7%), has-mir-1914-5p (96%), has-mir-1233-5p (97.7%), and has-mir-1290(95.4%). In **Supplementary Figure S2**, we used a boxplot to display the expression levels of these 10 candidate miRNAs in the cancer and non-cancer groups. In the boxplots, it is clear that four of the miRNAs has-miR-1233-5p, has-miR-1914-5p, has-miR-4675, and has-miR-5100 have higher expression level with various cut-off for cancerous samples, and on average, four of them (has-miR-1228-5p, has-miR-3184-5p, has-miR-6784-5p, and has-miR-6800-5p) have lower expression level for cancerous samples. We used heatmap plots by implementing the “heatmaply” package to underpin the potential relationships between features and the hierarchical clustering

TABLE 1 Hyperparameters and predictive power of models for ovarian cancer classification.

Classifier	Hyperparameters	AUC <sup>a</sup> (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Negative predictive value (NPV) %	Positive predictive value (PPV) %
Logistic regression	Parameters <sup>b</sup>	99.77	100.0	100	100.0	100.0	100.0
Decision trees	Cp = 0.01014493 <sup>c</sup>	98.30	91.30	97.41	97.10	88.10	94.0
Random forest	Mtry = 2 <sup>d</sup>	100.0	96.74	99.55	100.0	94.55	100.0
Artificial neural network	Size = 3 <sup>e</sup> and decay = 0.1 <sup>f</sup>	99.93	100.0	98.84	98.74	100.0	100.0
XGBoosting	nrounds = 50, max_depth = 2, eta = 0.4 <sup>g</sup> gamma = 0 <sup>h</sup> colsample_bytree <sup>i</sup> = 0.6 min_child_weight <sup>j</sup> = 1 and subsample = 0.75 <sup>k</sup>	99.99	98.91	99.28	100.0	100.0	98.11

<sup>a</sup>The area under the receiver operating characteristic curve (maximum) was used to select the optimal model.

<sup>b</sup>The formula for logistic regression for the prediction of ovarian cancer is

$$p = \left( 1 + e^{-\left( -10.463 - 18.25(\text{has.miR.5100}) - 29.63(\text{has.miR.6800.5p}) - 9.30(\text{has.miR.6784.5p}) - 7.38(\text{has.miR.3184.5p}) + 2.702(\text{has.miR.1228.5p}) + 11.33(\text{has.miR.4675}) - 8.19(\text{has.miR.1469}) + 0(\text{has.miR.1914.5p}) + 5.70(\text{has.miR.1233.5p}) + 9.08(\text{has.miR.1290}) \right)} \right)^{-1}$$

<sup>c</sup>The complexity parameter (cp) is used to control the size of the decision tree and to select the optimal tree size. If the cost of adding an additional variable to the decision tree from the current node is above the value of the cp, then tree building does not continue.

<sup>d</sup>mtry is the number of variables available for splitting at each tree node. In the random forests literature, this is referred to as the mtry parameter.

<sup>e</sup>Size is the number of units in a hidden layer.

<sup>f</sup>Decay is the regularization parameter used to avoid over-fitting.

<sup>g</sup>max-depth is used to control over-fitting as higher depth will allow the model to learn relations very specific to a particular sample.

<sup>h</sup>gamma A node is split only when the resulting split gives a positive reduction in the loss function. Gamma specifies the minimum loss reduction required to make a split, which makes the algorithm conservative. The values can vary depending on the loss function and should be tuned.

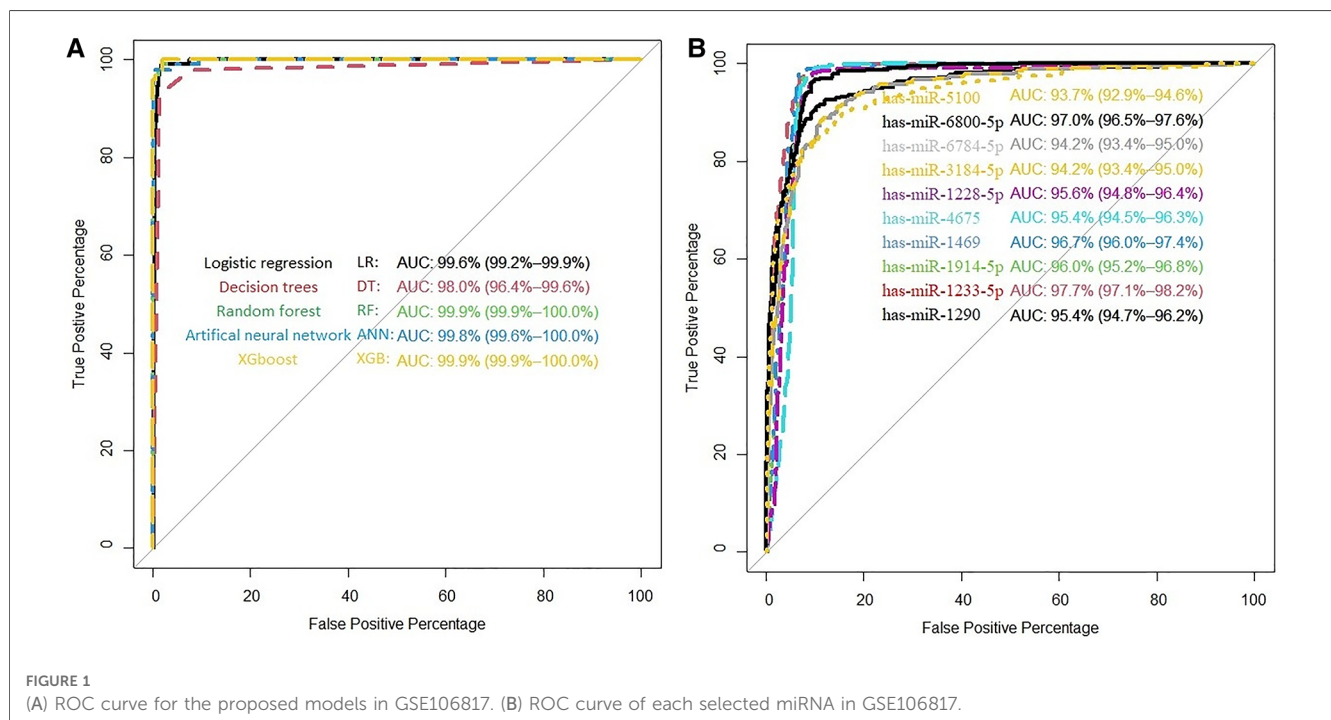
<sup>i</sup>Denotes the fraction of columns to be randomly sampled for each tree.

<sup>j</sup>min\_child\_weight is used to control over-fitting. Higher values prevent a model from learning relations that might be highly specific to the particular sample selected for a tree. Too high values can lead to under-fitting; hence, it should be tuned using CV.

<sup>k</sup>subsample lower values make the algorithm more conservative and prevent over-fitting but too small values might lead to under-fitting.

analysis using the selected features to recognize different samples in the internal discovery data sets. **Supplementary Figure S3** shows a promising result of the hierarchical clustering analysis (heatmap) using the 10 identified miRNAs to differentiate between

cancerous and non-cancerous samples in GSE106817. The selected microRNAs are differently expressed in the non-cancer and cancerous classes. This is well illustrated by drawing the heatmap (**Supplementary Figure S3**).



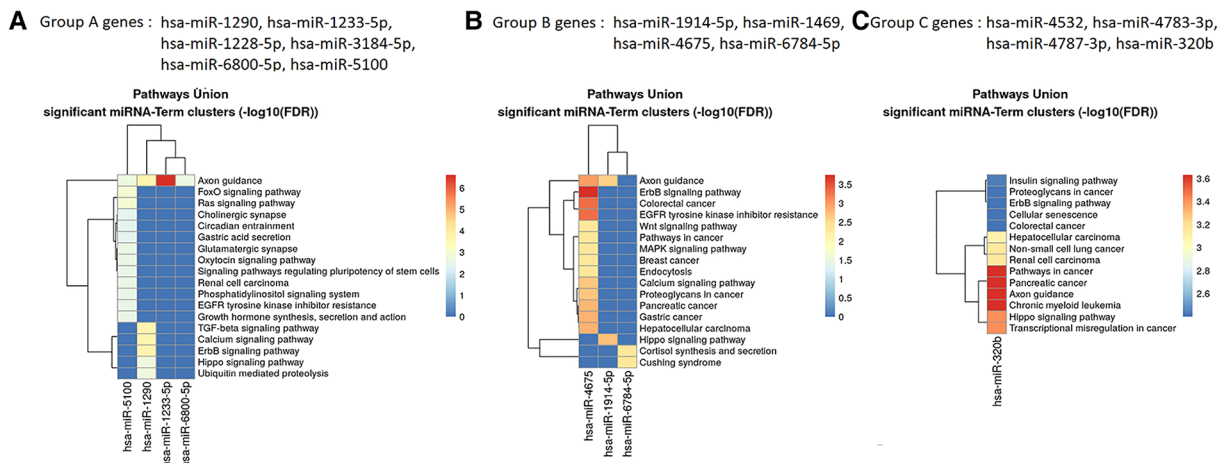


FIGURE 2 Targeted pathway clusters/heatmap presenting the top 10 Kyoto Encyclopedia of Genes and Genomes pathways regulated by the miRNAs ( $P < 0.005$ ; DIANA/miRPath v.4).

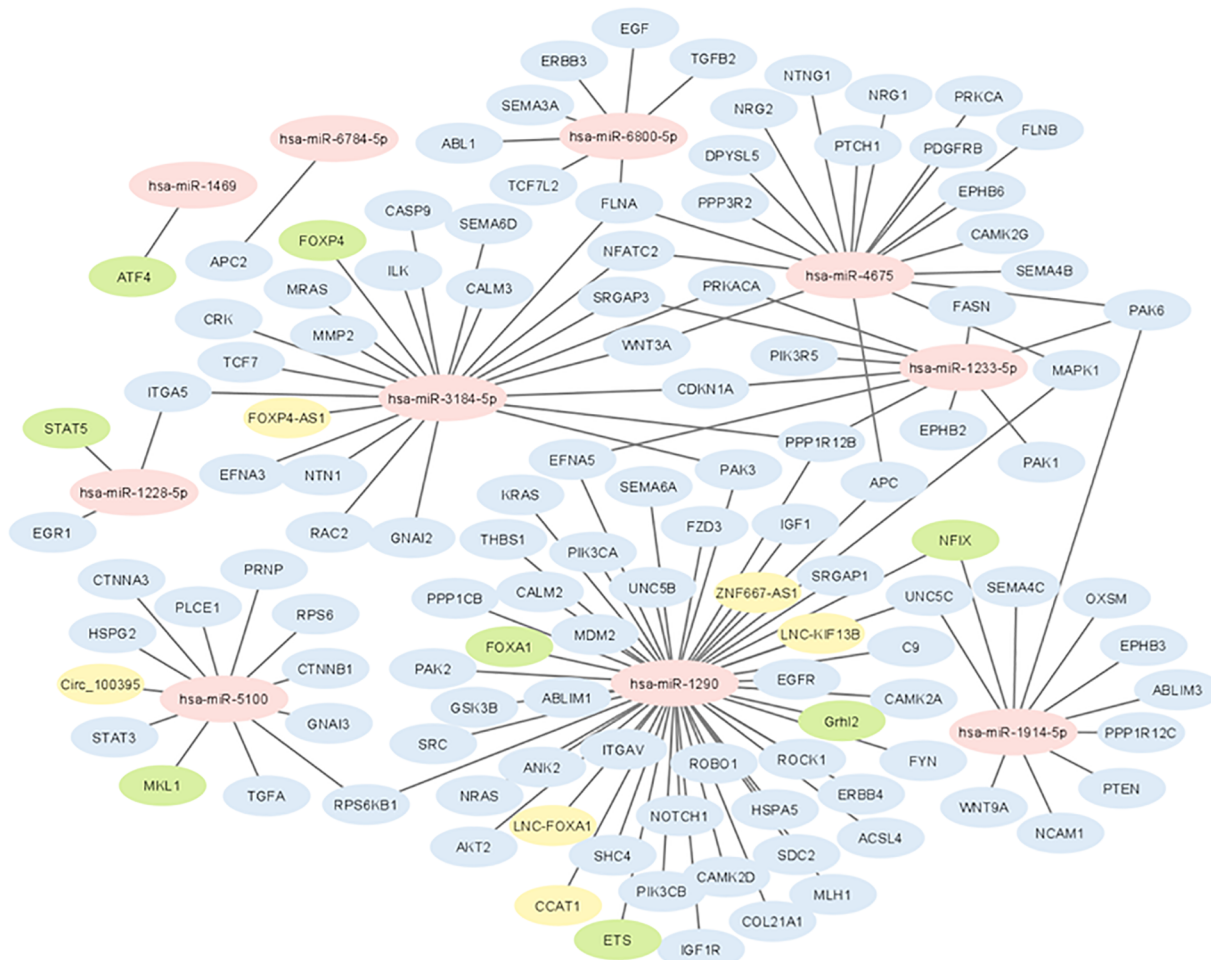


FIGURE 3 Network of interactions between selected miRNAs with coding genes and long non-coding RNAs. Yellow colored genes represent LNC-RNAs and green colored genes represent transcription factors.

### 3.2. External validation data sets

**Supplementary Table S1B,C** demonstrate the performance of miRNAs individually in proposed classification models in external validation data sets, as seen by the fact that almost all miRNAs have higher AUC in GSE113486 than in GSE113740. But in detail of each data set, **Supplementary Table S3** shows the results of the external data set models based on the Boruta feature selection algorithm. As seen in **Supplementary Table S3B** LR and ANN have a value of 100 in all seven criteria and XGB has an AUC, specificity, and PPV of 100. **Supplementary Figure S4A** shows that all models for GSE113486 yielded 100% AUC except DT and RF. **Supplementary Figure S4B** illustrates how biomarkers perform individually has-mir-5100 (95.8%), has-mir-6800-5p (99.7%), has-mir-6784-5p (97.5%), has-mir-3184-5p (93.9%), has-mir-1228-5p (99.8%), has-mir-4675 (99.1%), has-mir-1469 (100%), has-mir-1914-5p (99.8%), has-mir-1233-5p (99.4%), and has-mir-1290 (96.4%) in GSE113486. Boxplots show us that six of miRNAs (has-mir-1290, has-mir-1233-5p, has-mir-1914-5p, has-mir-1469, has-mir-4675, and has-mir-5100) have upregulated to ovarian cancer samples in GSE113486 (**Supplementary Figure S5**). In GSE113740, as the second external validation data set, we can see the result of AUC of LR, RF, ANN, and XGB over 94% in **Supplementary Figure S6A**. We also found AUC for these 10 miRNAs (individually) in external data sets that included individually has-mir-5100 (90.6%), has-mir-6800-5p (89.7%), has-mir-6784-5p (74.4%), has-mir-3184-5p (74.4%), has-mir-1228-5p (85.2%), has-mir-4675 (79.7%), has-mir-1469 (84.4%), has-mir-1914-5p (81.5%), has-mir-1233-5p (86.5%), and has-mir-1290 (91%) as shown in **Supplementary Figure S6B**. **Supplementary Table S3C** shows us that RF and XGB have the highest value in Kappa (72.96 and 71.96) in AUC and accuracy (97.2, 93.75), as seen ANN has 100 of sensitivity and NPV. Boxplots (**Supplementary Figure S7**) show us that six of miRNAs (has-mir-1290, has-mir-1233-5p, has-mir-1914-5p, has-mir-1469, has-mir-4675, and has-mir-5100) have high expression level in ovarian cancer samples. **Supplementary Figures S8, S9** show a promising result of the hierarchical clustering analysis (heatmap) using the 10 identified miRNAs to differentiate between the cancerous and non-cancerous samples in GSE113486 and GSE113740, respectively.

## 4. Discussion

It is critical to find and develop non-invasive, sensitive, and specific biomarkers to identify ovarian cancer in its early stages to effectively manage ovarian cancer patients. Fortunately, despite these limitations, newly discovered small RNAs called microRNAs have the potential to serve as effective non-invasive biomarkers for ovarian cancer (61, 62). Therefore, in this study, we used effective strategies and identified 10 miRNAs, hsa-miR-5100, hsa-miR-6800-5p, hsa-miR-6784-5p, hsa-miR-3184-5p, hsa-miR-1228-5p, hsa-miR-4675, hsa-miR-1469, hsa-miR-1914-5p, hsa-miR-1233-5p, and hsa-miR-1290, as strong potential biomarkers for ovarian cancer.

TABLE 2 Summary of the role of selected miRNAs in cancer.

miRNA	Cancer type	Reference
hsa-miR-1290	Lung	Zhang et al. (64)
hsa-miR-1290	Colorectal	Imaoka et al. (65) Ye et al. (66)
hsa-miR-1290	Hepatocellular	Wang et al. (67)
hsa-miR-1290	Advanced oral squamous cell carcinoma	Nakashima et al. (68)
hsa-miR-1290	Pancreatic	Wei et al. (69)
hsa-miR-1290	Ovarian	Kobayashi et al. (70) Li et al. (71)
hsa-miR-1233-5p	Renal cell carcinoma	Dias et al. (72)
hsa-miR-1914-5p	Colorectal	Liu et al. (73)
hsa-miR-1914-5p	Epithelial ovarian	Chong et al. (74)
hsa-miR-1469	Pancreatic	Shams et al. (75)
hsa-miR-1469	Laryngeal	Ma et al. (76)
hsa-miR-1469	Colon	Gungormez et al. (77)
hsa-miR-4675	Breast	Lai et al. (78)
hsa-miR-4675	Various types	Chen and Dhahbi (18)
hsa-miR-3184-5p	Breast	Rajaraman et al. (79)
hsa-miR-3184-5p	Ovarian	Alshamrani (80)
hsa-miR-3184-5p	Various types	Chen and Dhahbi (18)
hsa-miR-6800-5p	Epithelial ovarian	Tuncer et al. (81)
hsa-miR-5100	Various types	Chen and Dhahbi (18)
hsa-miR-5100	Epithelial ovarian	Tuncer et al. (81)
hsa-miR-5100	Pancreatic	Chijiwa et al. (82) Shams et al. (75)
hsa-miR-5100	Esophageal	Song et al. (83)
hsa-miR-1228-5p	Breast	Peña-Chilet et al. (84)
hsa-miR-1228-5p	Various types	Hu et al. (85)
hsa-miR-1228-5p	Breast	Cilek et al. (86)
hsa-miR-1228-5p	Hepatocellular	Morishita et al. (87)
hsa-miR-1228-5p	Epithelial ovarian	Chen et al. (88)
hsa-miR-1228-5p	Pancreatic	Wang et al. (89)
hsa-miR-6784-5p	Hepatocellular	Morishita et al. (87)
hsa-miR-6784-5p	Various types	Alshamrani (80)
hsa-miR-6784-5p	Esophageal	Song et al. (83)

### 4.1. Biological insight

The results of the biological insight section tell us about cell analysis for miRNAs that were found in this study based on the findings of the previous studies. The DIANA tool miRPath v.4 was used to perform the pathway enrichment analysis, based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. The target genes of miRNA were identified using TargetScan v8.0 databases. The settings of the software were *P*-value threshold = 0.005 and the FDR correction filter were ticked. It should be mentioned that we used two methods to find the target genes: the first one is the genes union and the second is the pathway union. To investigate the efficiency of the set of biomarkers selected by Boruta and their superiority over the previous similar work done by Hamidi et al. (21), three groups of miRNAs were analyzed by miRPath v.4: (A) common biomarkers of the current study and the previous study by Hamidi et al. (21); (B) biomarkers selected by Boruta in the present study and not identified in the previous work; and (C) biomarkers that were selected in the previous study and were not identified in the current study. The list of genes of these



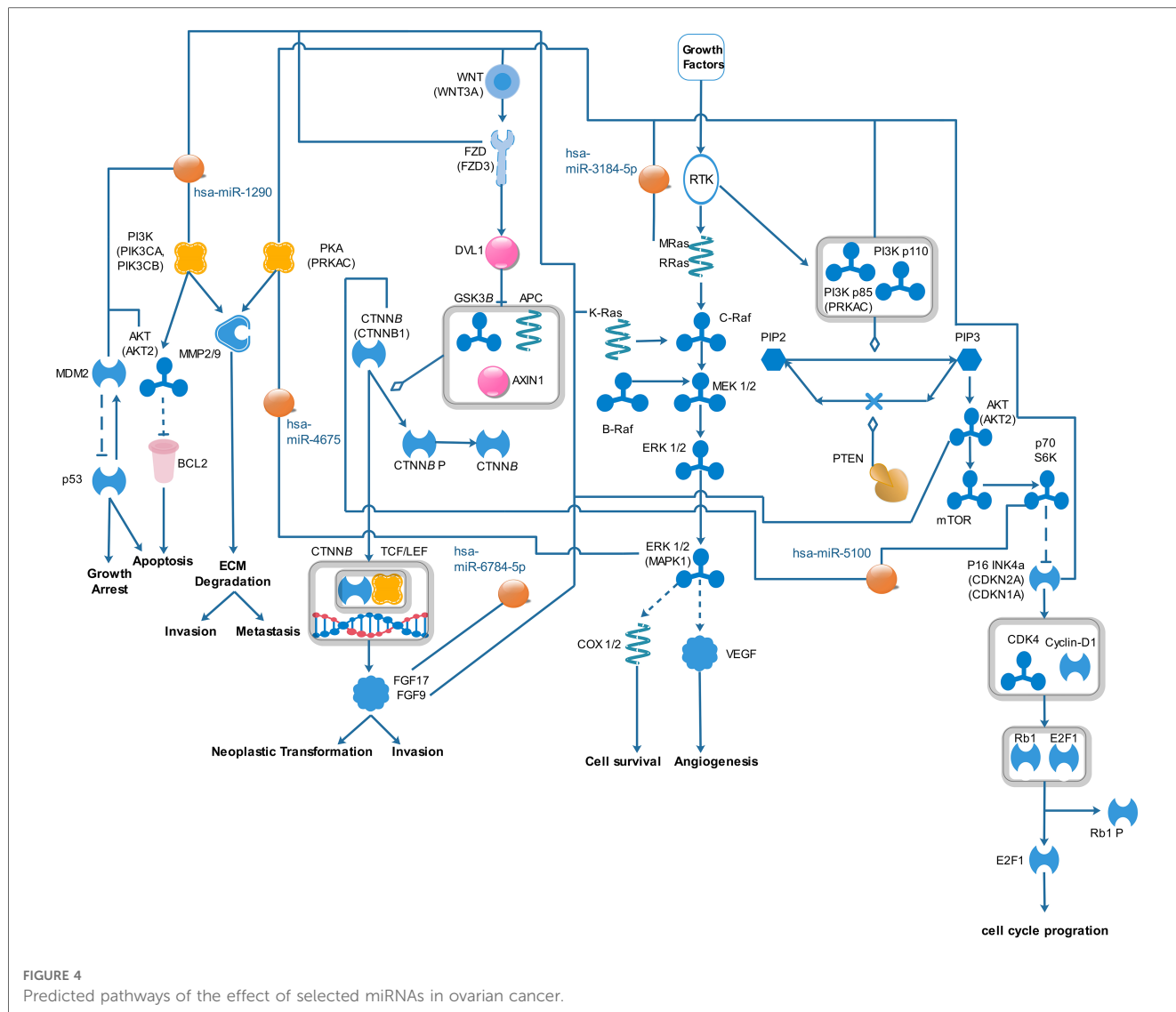
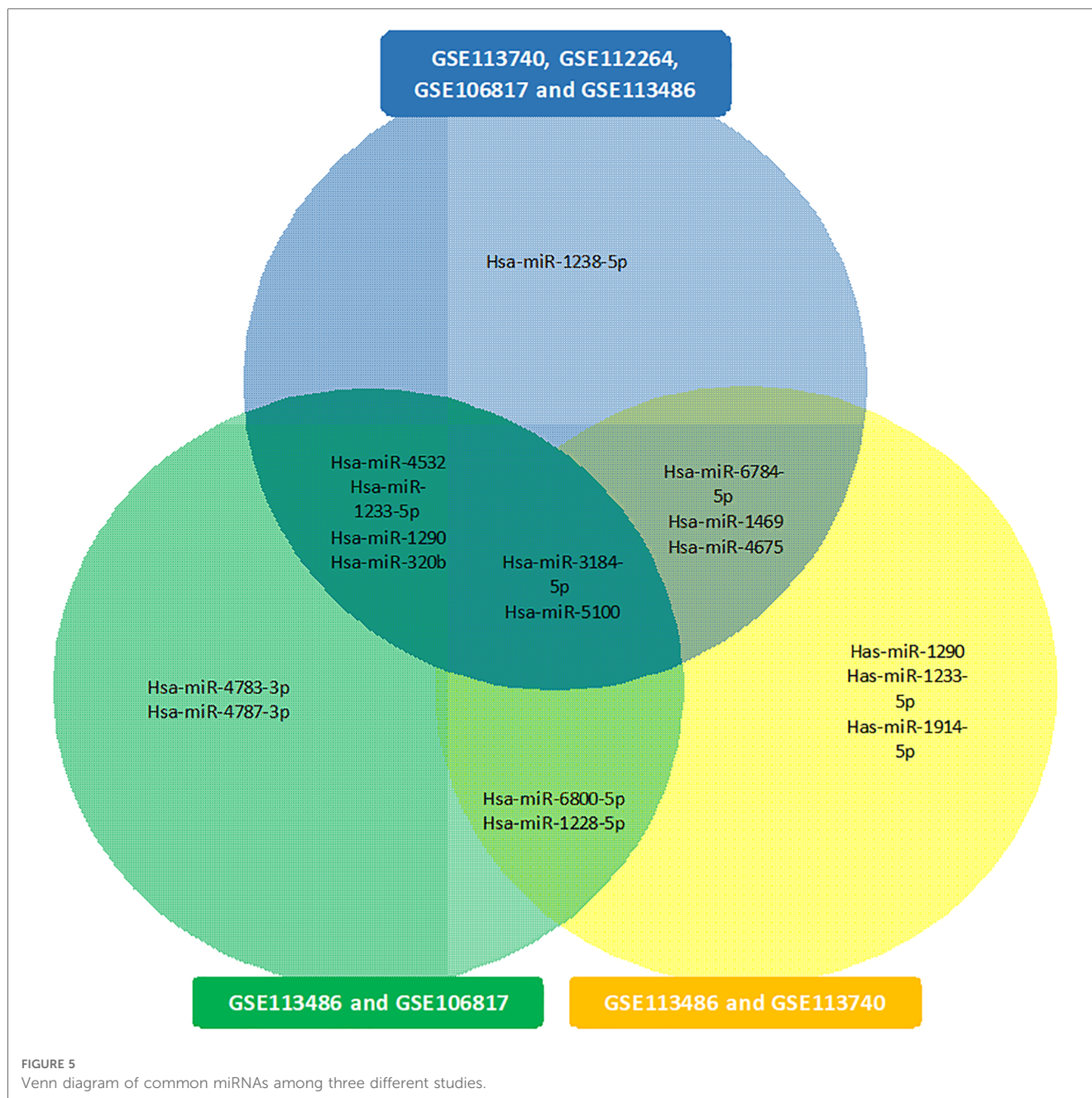


FIGURE 4 Predicted pathways of the effect of selected miRNAs in ovarian cancer.

three groups and their analysis results by miRPath v.4 tool are shown in **Figure 2**. As shown in **Figure 2A**, among the six common genes between the present and previous work, four genes are involved in at least one known cancer pathway (axon guidance). Among those four genes, hsa-miR-5100 and hsa-miR-1290 are involved in several well-known and important pathways in cancer. **Figure 2B** shows that among the four specific genes identified by the Boruta technique, three genes are involved in at least two well-known pathways in cancer, among which hsa-miR-4675 is involved in several pathways. However, in **Figure 2C**, among the four specific genes identified in the previous work of Hamidi et al. (21), only the hsa-miR-320b gene is involved in several important cancer pathways. It should be noted that there are six common paths between Groups A and B, while there are four common paths between A and C. This means that there are more correlation between genes of Group A and B than of Group A and C. This interpretation shows the biological superiority of Boruta's technique over the previous work. A comparison between routes of Group B and C also provides interesting results. Eight pathways are common between the two groups, which are proteoglycans in cancer, ErbB

signaling, colorectal cancer, hepatocellular cancer, pathways in cancer, pancreatic cancer, axon guidance, and Hippo signaling. Axon guidance pathway is common among all the three groups. Many axon guidance molecules regulate cell migration and apoptosis in normal and tumorigenic tissues (63). **Supplementary Table S4** shows the target genes of the selected microRNAs and the associated KEGG pathways from the genes union method, which indicates the significance of the relationship between the microRNAs and the corresponding pathways under the specified threshold values. **Figure 3** shows the network of miRNAs and identified target genes. In this figure, transcription factors and LNC-RNAs have also been added through some studies. References for these interactions are described in **Supplementary Table S4**.

In **Supplementary Table S5**, we only selected seven pathways because only the pathways that had very high correlation with miRNAs were selected (including a *P*-value of < 0.002). Among the top seven pathways identified, based on *P*-value, were pathways associated with fatty acid biosynthesis, prion diseases, axon guidance, glioma, ErbB signaling pathway, proteoglycans in cancer, and endometrial cancer. All signaling pathways related to miRNAs were



used from known pathways, and in general, they play an important role in all types of cancer. According to the KEGG database, some of the published articles confirm the role of some of the selected miRNAs in cancer directly. A number of these documents are summarized in **Table 2**. **Figure 4** shows the predicted pathways of the effect of some of the selected microRNAs that have been taken from the <https://targetexplorer.ingenuity.com/index.htm>.

**Figure 5** presents the common miRNAs between two related studies (18, 21) and miRNAs that were obtained in our study. There is some evidence in the literature for the biomarkers included in our study. Hamidi et al. (21) showed that hsa-miR-5100, hsa-miR-1233-5p, hsa-miR-4532, hsa-miR-1290, has-miR-3184-5p, and hsa-miR-320b could potentially be employed as important biomarkers in ovarian cancer. Jeon et al. (17)

investigated that miRNA-1290 in the epithelial ovarian cancer group was significantly overexpressed in serum exosomes and tissues as compared with the benign ovarian neoplasm group. Ying et al. (90) expressed that microarray data analysis showed that hsa-miR-1290 was differentially expressed between COC1 (DDP-sensitive) and COC1/DDP (DDP-resistant) tumor cell lines. Chen et al. (18) showed that only five balanced miRNAs were determined to be important in cancer diagnosis: hsa-miR-663a, hsa-miR-6802-5p, hsa-miR-6784-5p, hsa-miR-3184-5p, and hsa-miR-8073. Furthermore, Chen et al. (18) found that hsa-miR-3184-5p can act as an early biomarker of bladder cancer and as a key regulator of breast cancer. Also, hsa-miR-6784-5p has been reported to be a sensitive serum biomarker for ovarian cancer diagnosis and a key regulator for breast cancer.

In the end, we note that although there are fundamental differences between microarray and RNA-Seq methods for obtaining gene expression data, the data matrix obtained from both methods is completely similar after performing the necessary pre-processing. Therefore, our method is also applicable to RNA-Seq data.

## 5. Strengths and limitations

This study provides several advantages. Firstly, to identify the relevant and important miRNAs, we utilized a robust variable selection method and a novel random forest-based feature selection of a machine-learning approach to identify and select the relevant and important miRNAs for ovarian cancer diagnosis, using Boruta as a novel random forest-based feature selection in the machine-learning techniques that has known roles in dimension reduction and select properties variables. Secondly, we used logistic regression and four of the most used machine-learning methods to predict and classify ovarian cancer. Thirdly, we selected three GEO data sets and ensured that they were from a similar platform, and used them in the evaluation stages. The first limitation of this study is that the biomarkers obtained in this study for ovarian cancer were not compared with the other common types of cancer in females. Secondly, the result of this study is possibly appropriate for a specific race or area because of the main data set.

## 6. Conclusion

Our study aimed to investigate reliable classification biomarkers in ovarian cancer. After utilizing Boruta for identifying the important biomarkers, we found 10 miRNAs that have high reliability in evaluating output from each classification model. The Hsa-miR-5100, hsa-miR-6800-5p, hsa-miR-6784-5p, hsa-miR-3184-5p, hsa-miR-1228-5p, hsa-miR-4675, hsa-miR-1469, hsa-miR-1914-5p, hsa-miR-1233-5p, and hsa-miR-1290 had significant differential expression in all models, especially in the two data sets studied (GSE106817, GSE113486). Except for decision trees, all the proposed models have performed fairly well in terms of the detection accuracy for ovarian cancer in the validation data sets. The LR, RF, ANN, and XGB in GSE106817 and GSE113486 data sets had over 99% AUC, and in GSE113740 over 94%. Even though this study presented some additional biomarkers for possible consideration in future research, the analyses in these data sets do not support the immediate clinical use of these biomarkers without more rigorous testing in large case-control and cohort studies.

## Data availability statement

The original contributions presented in the study are included in the article/**Supplementary Material**; further inquiries can be directed to the corresponding author.

## Ethics statement

The studies involving human participants were reviewed and approved by the ethics committee of Tabriz University of Medical Sciences (IR.TBZMED.REC.1400.006). The patients/participants provided their written informed consent to participate in this study.

## Author contributions

RA, NG, and FH contributed to the conception and design of the study. RA, NG, and FH performed the statistical analysis. FH wrote the first draft of the manuscript. HY, EB and FH wrote the biological discussion section. RA, FH, PS and JM wrote sections of the manuscript. All authors contributed to the manuscript revision and read and approved the submitted version.

## Acknowledgments

The authors would like to thank all those who spent their valuable time participating in this research project, and we are also immensely grateful to the “anonymous” reviewers.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fdgth.2023.1187578/full#supplementary-material>

## References

- Lheureux S, Gourley C, Vergote I, Oza AM. Epithelial ovarian cancer. *Lancet*. (2019) 393(10177):1240–53. doi: 10.1016/S0140-6736(18)32552-2
- Reid BM, Permeth JB, Sellers TA. Epidemiology of ovarian cancer: a review. *Cancer Biol Med*. (2017) 14(1):9. doi: 10.20892/j.issn.2095-3941.2016.0084
- Cabasag CJ, Fagan PJ, Ferlay J, Vignat J, Laversanne M, Liu L, et al. Ovarian cancer today and tomorrow: a global assessment by world region and human development index using GLOBOCAN 2020. *Int J Cancer*. (2022) 151(9):1535–41. doi: 10.1002/ijc.34002
- Miller KD, Siegel RL, Lin CC, Mariotto AB, Kramer JL, Rowland JH, et al. Cancer treatment and survivorship statistics, 2016. *CA Cancer J Clin*. (2016) 66(4):271–89. doi: 10.3322/caac.21349
- Carioli G, Bertuccio P, Boffetta P, Levi F, La Vecchia C, Negri E, et al. European cancer mortality predictions for the year 2020 with a focus on prostate cancer. *Ann Oncol*. (2020) 31(5):650–8. doi: 10.1016/j.annonc.2020.02.009
- Iorio MV, Visone R, Di Leva G, Donati V, Petrocca F, Casalini P, et al. MicroRNA signatures in human ovarian cancer. *Cancer Res*. (2007) 67(18):8699–707. doi: 10.1158/0008-5472.CAN-07-1936
- Du Bois A, Reuss A, Pujade-Lauraine E, Harter P, Ray-Coquard I, Pfisterer J. Role of surgical outcome as prognostic factor in advanced epithelial ovarian cancer: a combined exploratory analysis of 3 prospectively randomized phase 3 multicenter trials: by the arbeitgemeinschaft gynaekologische onkologie studien-gruppe ovarialkarzinom (AGO-OVAR) and the groupe d'investigateurs nationaux pour les etudes des cancers de l'Ovaire (GINECO). *Cancer*. (2009) 115(6):1234–44. doi: 10.1002/cncr.24149
- Zheng H, Zhang L, Zhao Y, Yang D, Song F, Wen Y, et al. Plasma miRNAs as diagnostic and prognostic biomarkers for ovarian cancer. *PLoS One*. (2013) 8(11):e77853. doi: 10.1371/journal.pone.0077853
- Bartels CL, Tsongalis GJ. MicroRNAs: novel biomarkers for human cancer. *Clin Chem*. (2009) 55(4):623–31. doi: 10.1373/clinchem.2008.112805
- Flavin R, Smyth P, Barrett C, Russell S, Wen H, Wei J, et al. miR-29b expression is associated with disease-free survival in patients with ovarian serous carcinoma. *Int J Gynecologic Cancer*. (2009) 19(4):641–7. doi: 10.1111/IGC.0b013e3181a48c9
- Schwarzenbach H, Nishida N, Calin GA, Pantel K. Clinical relevance of circulating cell-free microRNAs in cancer. *Nat Rev Clin Oncol*. (2014) 11(3):145–56. doi: 10.1038/nrclinonc.2014.5
- Yokoi A, Yoshioka Y, Hirakawa A, Yamamoto Y, Ishikawa M, Ikeda S-i, et al. A combination of circulating miRNAs for the early detection of ovarian cancer. *Oncotarget*. (2017) 8(52):89811. doi: 10.18632/oncotarget.20688
- Matsuzaki J, Ochiya T. Circulating microRNAs and extracellular vesicles as potential cancer biomarkers: a systematic review. *Int J Clin Oncol*. (2017) 22:413–20. doi: 10.1007/s10147-017-1104-3
- Hamidi F, Gilani N, Belaghi RA, Sarbakhsh P, Edgünlü T, Santaguida P. Exploration of potential miRNA biomarkers and prediction for ovarian cancer using artificial intelligence. *Front Genet*. (2021) 12:724785. doi: 10.3389/fgene.2021.724785
- Chung Y-W, Bae H-S, Song J-Y, Lee JK, Lee NW, Kim T, et al. Detection of microRNA as novel biomarkers of epithelial ovarian cancer from the serum of ovarian cancer patient. *Int J Gynecologic Cancer*. (2013) 23(4):673–9. doi: 10.1097/IGC.0b013e31828c166d
- Yuan F, Li Z, Chen L, Zeng T, Zhang Y-H, Ding S, et al. Identifying the signatures and rules of circulating extracellular microRNA for distinguishing cancer subtypes. *Front Genet*. (2021) 12:651610. doi: 10.3389/fgene.2021.651610
- Jeon H, Seo SM, Kim TW, Ryu J, Kong H, Jang SH, et al. Circulating exosomal miR-1290 for diagnosis of epithelial ovarian cancer. *Curr Issues Mol Biol*. (2022) 44(1):288–300. doi: 10.3390/cimb44010021
- Chen JW, Dhahbi J. Identification of four serum miRNAs as potential markers to screen for thirteen cancer types. *PLoS One*. (2022) 17(6):e0269554. doi: 10.1371/journal.pone.0269554
- Yaghoobi H, Babaei E, Hussen BM, Emami A. EBST: an evolutionary multi-objective optimization based tool for discovering potential biomarkers in ovarian cancer. *IEEE/ACM Trans Comput Biol Bioinform*. (2020) 18(6):2384–93. doi: 10.1109/TCBB.2020.2993150
- Zhang A, Hu H. A novel blood-based microRNA diagnostic model with high accuracy for multi-cancer early detection. *Cancers (Basel)*. (2022) 14(6):1450. doi: 10.3390/cancers14061450
- Hamidi F, Gilani N, Belaghi RA, Sarbakhsh P, Edgünlü T, Santaguida P. Exploration of potential miRNA biomarkers and prediction for ovarian cancer using artificial intelligence. *Front Genet*. (2021) 12:2079. doi: 10.3389/fgene.2021.724785
- Tripathi YM, Chatla SB, Chang Y-CI, Huang L-S, Shieh GS. A nonlinear correlation measure with applications to gene expression data. *PLoS One*. (2022) 17(6):e0270270. doi: 10.1371/journal.pone.0270270
- Nithya B, Ilango V. Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction. *SN Appl Sci*. (2019) 1:1–16. doi: 10.1007/s42452-019-0645-7
- Chen R-C, Dewi C, Huang S-W, Caraka RE. Selecting critical features for data classification based on machine learning methods. *J Big Data*. (2020) 7(1):52. doi: 10.1186/s40537-020-00327-4
- Ali NM, Aziz N, Besar R. Comparison of microarray breast cancer classification using support vector machine and logistic regression with LASSO and boruta feature selection. *Indones J Electr Eng Comput Sci*. (2020) 20(2):712–9. doi: 10.11591/ijeecs.v20.i2.pp712-719
- Fortino V, Kinaret P, Fyhrquist N, Alenius H, Greco D. A robust and accurate method for feature selection and prioritization from multi-class OMICs data. *PLoS One*. (2014) 9(9):e107801. doi: 10.1371/journal.pone.0107801
- Kursa MB, Rudnicki WR. Feature selection with the boruta package. *J Stat Softw*. (2010) 36:1–13. doi: 10.18637/jss.v036.i11
- Chen K-H, Wang K-J, Tsai M-L, Wang K-M, Adrian AM, Cheng W-C, et al. Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. *BMC Bioinform*. (2014) 15(1):1–10. doi: 10.1186/1471-2105-15-1
- Trivedi SK. A study on credit scoring modeling with different feature selection and machine learning approaches. *Technol Soc*. (2020) 63:101413. doi: 10.1016/j.techsoc.2020.101413
- Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. *Brief Bioinform*. (2019) 20(2):492–503. doi: 10.1093/bib/bbx124
- Acharjee A, Larkman J, Xu Y, Cardoso VR, Gkoutos GV. A random forest based biomarker discovery and power analysis framework for diagnostics research. *BMC Med Genomics*. (2020) 13(1):1–14. doi: 10.1186/s12920-020-00826-6
- Kursa MB. Robustness of random forest-based gene selection methods. *BMC Bioinform*. (2014) 15:1–8. doi: 10.1186/1471-2105-15-8
- Yokoi A, Matsuzaki J, Yamamoto Y, Yoneoka Y, Takahashi K, Shimizu H, et al. Integrated extracellular microRNA profiling for ovarian cancer screening. *Nat Commun*. (2018) 9(1):4319. doi: 10.1038/s41467-018-06434-4
- Usuba W, Urabe F, Yamamoto Y, Matsuzaki J, Sasaki H, Ichikawa M, et al. Circulating miRNA panels for specific and early detection in bladder cancer. *Cancer Sci*. (2019) 110(1):408–19. doi: 10.1111/cas.13856
- Yamamoto Y, Kondo S, Matsuzaki J, Esaki M, Okusaka T, Shimada K, et al. Highly sensitive circulating microRNA panel for accurate detection of hepatocellular carcinoma in patients with liver disease. *Hepatol Commun*. (2020) 4(2):284–97. doi: 10.1002/hep4.1451
- Wiemken TL, Kelley RR. Machine learning in epidemiology and health outcomes research. *Annu Rev Public Health*. (2019) 41:21–36. doi: 10.1146/annurev-publichealth-040119-094437
- Speiser JL, Miller ME, Tooze J, Ip E. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Syst Appl*. (2019) 134:93–101. doi: 10.1016/j.eswa.2019.05.028
- Lekchnov EA, Amelina EV, Bryzgunova OE, Zaporozhchenko IA, Konoshenko MY, Yarmoschuk SV, et al. Searching for the novel specific predictors of prostate cancer in urine: the analysis of 84 miRNA expression. *Int J Mol Sci*. (2018) 19(12):4088. doi: 10.3390/ijms19124088
- Hastie T, Tibshirani R, Friedman JH, Friedman JH. *The elements of statistical learning: Data mining, inference, and prediction*. Springer (2009).
- Huang J, Li Y-F, Xie M. An empirical analysis of data preprocessing for machine learning-based software cost estimation. *Inf Softw Technol*. (2015) 67:108–27. doi: 10.1016/j.infsof.2015.07.004
- Jović A, Brkić K, Bogunović N, editor. *A review of feature selection methods with applications. 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*; IEEE (2015).
- Stuart EA, Azur M, Frangakis C, Leaf P. Multiple imputation with large data sets: a case study of the children's mental health initiative. *Am J Epidemiol*. (2009) 169(9):1133–9. doi: 10.1093/aje/kwp026
- Azimi I, Pahikkala T, Rahmani AM, Niela-Vilén H, Axelin A, Liljeberg P. Missing data resilient decision-making for healthcare IoT through personalization: a case study on maternal health. *Future Gener Comput Syst*. (2019) 96:297–308. doi: 10.1016/j.future.2019.02.015
- Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. *J Big Data*. (2019) 6(1):1–54. doi: 10.1186/s40537-018-0162-3
- Sun Y, Wong AK, Kamel MS. Classification of imbalanced data: a review. *Int J Pattern Recognit Artif Intell*. (2009) 23(04):687–719. doi: 10.1142/S0218001409007326
- Fotouhi S, Asadi S, Kattan MW. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *J Biomed Inform*. (2019) 90:103089. doi: 10.1016/j.jbi.2018.12.003

47. Shanab AA, Khoshgoftaar TM, Wald R, Napolitano A, editor. *Impact of noise and data sampling on stability of feature ranking techniques for biological datasets*. 2012 IEEE 13th International Conference on Information Reuse & Integration (IRI); IEEE (2012).
48. Alpaydin E. *Introduction to machine learning*. MIT press (2020).
49. James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning*. Springer (2013).
50. Stoltzfus JC. Logistic regression: a brief primer. *Acad Emerg Med*. (2011) 18(10):1099–104. doi: 10.1111/j.1553-2712.2011.01185.x
51. Abdulqader QM. Applying the binary logistic regression analysis on the medical data. *Sci J Univ Zakho*. (2017) 5(4):330–4. doi: 10.25271/2017.5.4.388
52. Maimon OZ, Rokach L. *Data mining with decision trees: Theory and applications*. World Scientific (2014).
53. Qi Y. *Random forest for bioinformatics. Ensemble machine learning: Methods and applications*. Springer (2012). 307–23.
54. DeGregory K, Kuiper P, DeSilvio T, Pleuss J, Miller R, Roginski J, et al. A review of machine learning in obesity. *Obes Rev*. (2018) 19(5):668–85. doi: 10.1111/obr.12667
55. Klassen V, Safin A, Maltsev A, Andrianov N, Morozov S, Vladzmyrskyy A. AI-based screening of pulmonary tuberculosis: diagnostic accuracy. *J Ehealth Technol Appl*. (2018) 16(1):28–32.
56. Sherriff A, Ott J, Team AS. Artificial neural networks as statistical tools in epidemiological studies: analysis of risk factors for early infant wheeze. *Paediatr Perinat Epidemiol*. (2004) 18(6):456–63. doi: 10.1111/j.1365-3016.2004.00592.x
57. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, et al. Xgboost: extreme gradient boosting. R package version 04-2. *Journal of eHealth Technology and Application*. (2015) 16(1):1–4.
58. Kang C, Huo Y, Xin L, Tian B, Yu B. Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine. *J Theor Biol*. (2019) 463:77–91. doi: 10.1016/j.jtbi.2018.12.010
59. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. (2015) 162(1):55–63. doi: 10.7326/M14-0697
60. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform*. (2011) 12(1):1–8. doi: 10.1186/1471-2105-12-77
61. Pal MK, Jaiswar SP, Dwivedi VN, Tripathi AK, Dwivedi A, Sankhwar P. MicroRNA: a new and promising potential biomarker for diagnosis and prognosis of ovarian cancer. *Cancer Biol Med*. (2015) 12(4):328. doi: 10.7497/j.issn.2095-3941.2015.0024
62. Zhang B, Cai FF, Zhong XY. An overview of biomarkers for the ovarian cancer diagnosis. *Eur J Obstet Gynecol Reprod Biol*. (2011) 158(2):119–23. doi: 10.1016/j.ejogrb.2011.04.023
63. Chedotal A, Kerjan G, Moreau-Fauvarque C. The brain within the tumor: new roles for axon guidance molecules in cancers. *Cell Death Differ*. (2005) 12(8):1044–56. doi: 10.1038/sj.cdd.4401707
64. Zhang WC, Chin TM, Yang H, Nga ME, Lunny DP, Lim EKH, et al. Tumour-initiating cell-specific miR-1246 and miR-1290 expression converge to promote non-small cell lung cancer progression. *Nat Commun*. (2016) 7(1):11702. doi: 10.1038/ncomms11702
65. Imaoka H, Toiyama Y, Fujikawa H, Hiro J, Saigusa S, Tanaka K, et al. Circulating microRNA-1290 as a novel diagnostic and prognostic biomarker in human colorectal cancer. *Ann Oncol*. (2016) 27(10):1879–86. doi: 10.1093/annonc/mdw279
66. Ye L, Jiang T, Shao H, Zhong L, Wang Z, Liu Y, et al. miR-1290 is a biomarker in DNA-mismatch-repair-deficient colon cancer and promotes resistance to 5-fluorouracil by directly targeting hMSH2. *Mol Ther Nucl Acids*. (2017) 7:453–64. doi: 10.1016/j.omtn.2017.05.006
67. Wang Q, Wang G, Niu L, Zhao S, Li J, Zhang Z, et al. Exosomal MiR-1290 promotes angiogenesis of hepatocellular carcinoma via targeting SMEK1. *J Oncol*. (2021) 2021:6617700. doi: 10.1155/2021/6617700
68. Nakashima H, Yoshida R, Hirose A, Kawahara K, Sakata J, Arita H, et al. Circulating miRNA-1290 as a potential biomarker for response to chemoradiotherapy and prognosis of patients with advanced oral squamous cell carcinoma: a single-center retrospective study. *Tumor Biol*. (2019) 41(3):1010428319826853. doi: 10.1177/1010428319826853
69. Wei J, Yang L, Wu Y-n, Xu J. Serum miR-1290 and miR-1246 as potential diagnostic biomarkers of human pancreatic cancer. *J Cancer*. (2020) 11(6):1325. doi: 10.7150/jca.38048
70. Kobayashi M, Sawada K, Nakamura K, Yoshimura A, Miyamoto M, Shimizu A, et al. Exosomal miR-1290 is a potential biomarker of high-grade serous ovarian carcinoma and can discriminate patients from those with malignancies of other histological types. *J Ovarian Res*. (2018) 11:1–10. doi: 10.1186/s13048-018-0458-0
71. Li Y, Yao L, Liu F, Hong J, Chen L, Zhang B, et al. Characterization of microRNA expression in serous ovarian carcinoma. *Int J Mol Med*. (2014) 34(2):491–8. doi: 10.3892/ijmm.2014.1813
72. Dias F, Teixeira AL, Ferreira M, Adem B, Bastos N, Vieira J, et al. Plasmatic miR-210, miR-221 and miR-1233 profile: potential liquid biopsies candidates for renal cell carcinoma. *Oncotarget*. (2017) 8(61):103315. doi: 10.18632/oncotarget.21733
73. Liu S, Qu D, Li W, He C, Li S, Wu G, et al. Mir-647 and miR-1914 promote cancer progression equivalently by downregulating nuclear factor IX in colorectal cancer. *Mol Med Rep*. (2017) 16(6):8189–99. doi: 10.3892/mmr.2017.7675
74. Chong GO, Jeon H-S, Han HS, Son JW, Lee YH, Hong DG, et al. Differential microRNA expression profiles in primary and recurrent epithelial ovarian cancer. *Anticancer Res*. (2015) 35(5):2611–7. doi: 10.3892/ijmm.2014.1813
75. Shams R, Saberi S, Zali M, Sadeghi A, Ghafouri-Fard S, Aghdaei HA. Identification of potential microRNA panels for pancreatic cancer diagnosis using microarray datasets and bioinformatics methods. *Sci Rep*. (2020) 10(1):7559. doi: 10.1038/s41598-020-64569-1
76. Ma C-H, Zhang Y-X, Tang L-H, Yang X-J, Cui W-M, Han C-C, et al. MicroRNA-1469, a p53-responsive microRNA promotes genistein induced apoptosis by targeting Mcl1 in human laryngeal cancer cells. *Biomed Pharmacother*. (2018) 106:665–71. doi: 10.1016/j.biopha.2018.07.005
77. Gungormez C, Gumushan Aktas H, Dilsiz N, Borazan E. Novel miRNAs as potential biomarkers in stage II colon cancer: microarray analysis. *Mol Biol Rep*. (2019) 46:4175–83. doi: 10.1007/s11033-019-04868-7
78. Lai J, Wang H, Pan Z, Su F. A novel six-microRNA-based model to improve prognosis prediction of breast cancer. *Aging (Albany NY)*. (2019) 11(2):649. doi: 10.18632/aging.101767
79. Rajarajan D, Selvarajan S, Charan Raja MR, Kar Mahapatra S, Kasiappan R. Genome-wide analysis reveals miR-3184-5p and miR-181c-3p as a critical regulator for adipocytes-associated breast cancer. *J Cell Physiol*. (2019) 234(10):17959–74. doi: 10.1002/jcp.28428
80. Alshamrani AA. Roles of microRNAs in ovarian cancer tumorigenesis: two decades later, what have we learned? *Front Oncol*. (2020) 10:1084. doi: 10.3389/fonc.2020.10184
81. Tuncer SB, Erdogan OS, Erciyas SK, Saral MA, Celik B, Odemis DA, et al. miRNA expression profile changes in the peripheral blood of monozygotic discordant twins for epithelial ovarian carcinoma: potential new biomarkers for early diagnosis and prognosis of ovarian carcinoma. *J Ovarian Res*. (2020) 13(1):1–15. doi: 10.1186/s13048-020-00706-8
82. Chijiwa Y, Moriyama T, Ohuchida K, Nabae T, Ohtsuka T, Miyasaka Y, et al. Overexpression of microRNA-5100 decreases the aggressive phenotype of pancreatic cancer cells by targeting PODXL. *Int J Oncol*. (2016) 48(4):1688–700. doi: 10.3892/ijo.2016.3389
83. Song Y, Zhu S, Zhang N, Cheng L. Blood circulating miRNA pairs as a robust signature for early detection of esophageal cancer. *Front Oncol*. (2021) 11:723779. doi: 10.3389/fonc.2021.723779
84. Peña-Chilet M, Martínez MT, Pérez-Fidalgo JA, Peiró-Chova L, Oltra SS, Tormo E, et al. MicroRNA profile in very young women with breast cancer. *BMC Cancer*. (2014) 14(1):1–14. doi: 10.1186/1471-2407-14-529
85. Hu J, Wang Z, Liao BY, Yu L, Gao X, Lu S, et al. Human miR-1228 as a stable endogenous control for the quantification of circulating microRNAs in cancer patients. *Int J Cancer*. (2014) 135(5):1187–94. doi: 10.1002/ijc.28757
86. Ruggles DR, Freyman RL, Oxenham AJ. Influence of musical training on understanding voiced and whispered speech in noise. *PLoS One*. (2014) 9(1):e86980. doi: 10.1371/journal.pone.0086980
87. Morishita A, Iwama H, Fujihara S, Sakamoto T, Fujita K, Tani J, et al. MicroRNA profiles in various hepatocellular carcinoma cell lines. *Oncol Lett*. (2016) 12(3):1687–92. doi: 10.3892/ol.2016.4853
88. Chen X, Paranjape T, Stahlhut C, McVeigh T, Keane F, Nallur S, et al. Targeted resequencing of the microRNAome and 3' UTRome reveals functional germline DNA variants with altered prevalence in epithelial ovarian cancer. *Oncogene*. (2015) 34(16):2125–37. doi: 10.1038/onc.2014.117
89. Wang J, Raimondo M, Guha S, Chen J, Diao L, Dong X, et al. Circulating microRNAs in pancreatic juice as candidate biomarkers of pancreatic cancer. *J Cancer*. (2014) 5(8):696. doi: 10.7150/jca.10094
90. Ying H, Xu H, Lv J, Ying T, Yang Q. MicroRNA signatures of platinum-resistance in ovarian cancer. *Eur J Gynaecol Oncol*. (2015) 36(1):16–20. doi: 10.12892/ejgo2511.2015