Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Prediction of ovarian cancer using artificial intelligence tools

Seyed Mohammad Ayyoubzadeh
Tehran University of Medical Sciences
Marjan Ahmadi
Tehran University of Medical Sciences
Alireza Banaye Yazdipour
Tehran University of Medical Sciences
Fatemeh Ghorbani-Bidkorpeh
Shahid Beheshti University of Medical Sciences
Mahnaz Ahmadi (✉ imahnazahmadi@gmail.com)
Shahid Beheshti University of Medical Sciences

Research Article

Keywords: Artificial Intelligence, Machine Learning, Ovarian Cancer, Random Forest, Decision Tree, Support Vector Machine, Artificial Neural Network

Posted Date: December 12th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3728713/v1

License: 🅭 🅯 This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Purpose

Ovarian cancer is a common type of cancer and a leading cause of death in women. Therefore, accurate and fast prediction of ovarian tumors is crucial. One of the appropriate and precise methods for predicting and diagnosing this cancer is to build a model based on artificial intelligence methods. These methods provide a tool for predicting ovarian cancer according to the characteristics and conditions of each person.

Method

In this study, a dataset included records related to 171 cases of benign ovarian tumors and 178 records related to cases of ovarian cancer were analyzed. The dataset contains the records of blood test results and tumor markers of the patients. After data pre-processing, including removing outliers and replacing missing values, the weight of the effective factors was determined using information gain indices and the Gini index. In the next step, predictive models were created using Decision Trees, Support Vector Machine, Random Forest, and Artificial Neural Network models. The performance of these models was evaluated using the 10-fold cross-validation method using the indicators of accuracy, sensitivity, specificity, and the area under the ROC (Receiver operating characteristic) curve. Finally, by comparing the performance of the models, the best predictive model of ovarian cancer was selected.

Results

The most important predictive factors were HE4, CA125, and NEU. The Random Forest model was identified as the best predictive model with an accuracy of more than 86%. The predictive accuracy of Decision Tree, Support Vector Machine, and Artificial Neural Network models was estimated as 82.91%, 85.25%, and 79.35%, respectively. Various AI tools can be used with high accuracy and sensitivity in predicting ovarian cancer.

Conclusion

Therefore, the use of these tools can help specialists and patients with early, easier, and less expensive diagnosis of ovarian cancer. Future research can use AI by combining image data with serum biological indicators to develop new models and promote the diagnosis and treatment of ovarian cancer.

Introduction

Cancer is a malignancy characterized by high aggressiveness, low survival rates, and prolonged and costly treatment procedures. The disease's high recurrence and mortality rates make it imperative to achieve early detection and precise prognostication of cancer, as these measures are crucial for improving the likelihood of patient survival [1–3]. Ovarian cancer is one of the most prevalent forms of cancer affecting women. Every year, over 240,000 new cases of ovarian cancer are identified, and approximately 150,000 women lose their lives to this disease. Ovarian cancer consists of a diverse group of tumors that are categorized based on distinct histopathological and molecular characteristics. The most common form of ovarian cancer is epithelial ovarian cancer (EOC), which can be further divided into four main subtypes based on the appearance of tumor cells: serous, endometrioid, clear cell, and mucinous. The significant morbidity and mortality associated with ovarian cancer can be attributed to the late detection of the disease and reduced effectiveness of surgical or pharmacological treatments. Because symptoms tend to appear late and lack specificity, up to 75% of ovarian cancer cases are diagnosed at an advanced stage, and only around 20% of these individuals will survive for five years from the time of diagnosis [4, 5].

Different screening techniques like pelvic exams, transvaginal ultrasounds, CA125 cancer antigen tests, and MRI imaging are used to identify this disease. However, using any of these methods may not guarantee accurate diagnosis. For instance, pelvic examination and ultrasound have low sensitivity and specificity, while CA125 marker levels may not rise in all patients with ovarian cancer. Furthermore, an expert specialist is required for accurate diagnosis through MRI imaging, which can be challenging. Additionally, there is no proof of cost-effectiveness associated with any of these diagnostic methods [6–8]. The development of predictive tools has enabled patients and medical practitioners to carry out diagnostic procedures more accurately and quickly, while also enabling them to devise treatment plans that are well-suited to the specific needs of each patient. Artificial Intelligence (AI) systems have gained widespread adoption as a result of their numerous benefits, and can be employed to surmount the shortcomings of traditional diagnostic techniques. These systems have several advantages, such as their ability to handle large quantities of data, address instances of missing data, and adapt to new data inputs [9, 10].

AI techniques have been increasingly utilized in recent times for precise diagnostic applications across diverse disease categories. In recent years, various AI tools, especially machine learning and deep learning, have become popular for diagnosing and predicting various diseases, especially cancer, due to their advantages. For this reason, many studies have been published in this field [1, 7]. In addition, limited studies have been conducted concerning the prediction of ovarian cancer employing AI (machine learning) tools. However, due to the restrictions of these studies, the need for newer and more complete studies is felt [11, 12].

Therefore, this study proposes the adoption of artificial intelligence-based systems as prediction tools for ovarian cancer. In this regard, a set of data will be extracted from a dataset including the information of different patients, and AI methodologies will be employed to construct diversified models that can effectively predict ovarian cancer. The best-performing model will then be identified through subsequent evaluations.

Methods

Dataset

This study was conducted in 2022–2023. The data was collected from the online public repository, which includes the data of 349 patients with 49 characteristics as input (Table 1)

Table 1
Dataset description

Feature name	Type	Range	Missing Values
AFP	real	=[0.610–508]	24
AG	real	=[6.200–33.330]	1
Age	integer	=[15–83]	0
ALB	real	=[22–51.500]	10
ALP	integer	=[26–763]	10
ALT	integer	=[4–86]	10
AST	integer	=[7–78]	10
BASO#	real	=[0–0.120]	0
BASO%	real	=[0–1.940]	0
BUN	real	=[1.120–10.190]	0
Ca	real	=[0.920–2.830]	0
CA125	real	=[3.750–4468]	19
CA19-9	real	=[0.600–566.100]	34
CA72-4	real	=[0.200–158.500]	240
CEA	real	=[0.200–138.800]	22
CL	real	=[84.600–109.400]	0
CO2CP	real	=[16.200–34.300]	1
CREA	real	=[38.200–114]	0
DBIL	real	=[0.900–12.100]	10
EO#	real	=[0–0.400]	0
EO%	real	=[0–7.600]	0
GGT	integer	=[4–176]	10
GLO	real	=[14.100–47.600]	10
GLU.	real	=[3.570–12.440]	0
HCT	real	=[0.224–0.569]	0
HE4	real	=[16.710–3537.600]	20
HGB	real	=[61.800–189]	0

Feature name	Type	Range	Missing Values
IBIL	real	=[1–28.400]	10
K	real	=[3.080–5.400]	0
LYM#	real	=[0.350–3.490]	0
LYM%	real	=[3.900–51.600]	0
MCH	real	=[17.700–36.800]	0
MCV	real	=[61–107.900]	0
Menopause	binominal	=[0, 1]	0
Mg	real	=[0.650–1.370]	0
MONO#	real	=[0.070–0.970]	0
MONO%	real	=[0.300–21.300]	0
MPV	real	=[5.060–14.500]	2
Na	real	=[125.100–150.700]	0
NEU	real	=[37.200–92]	91
PCT	real	=[0.070–0.690]	2
PDW	real	=[8.800–22.800]	2
PHOS	real	=[0.570–1.750]	0
PLT	integer	=[74–868]	0
RBC	real	=[2.620–6.740]	0
RDW	real	=[10.920–22.200]	0
TBIL	real	=[2.500–38.300]	10
TP	real	=[32.900–86.800]	10
UA	real	=[96–632]	0
TYPE	binominal	=[0, 1]	0

Data analysis

The overall research steps are illustrated in Fig. 1. The data analysis methodology involved the following steps:

1. Data preprocessing: The RapidMiner version 9.10 software was used to clean the data by replacing missing values, removing outliers, and normalization of the data. This step was crucial to ensure the

accuracy of the subsequent analyses.

2. Factor weight determination: The weight of factors affecting ovarian cancer was determined using Information Gain and Gini Index methods. These methods helped to identify the most important factors that contribute to the development of ovarian cancer.
3. Modeling: Artificial intelligence models were created using classification techniques such as Decision Tree, Support Vector Machine, and Random Forest. The efficiency of the models was estimated using the indicators of accuracy, sensitivity, specificity, and area under the ROC curve. The models were evaluated using 10-Fold cross validation by accuracy, sensitivity, specificity, and ROC AUC indexes. The best model was selected based on its efficiency. The implemented blocks in RapidMiner studio are presented in Fig. 2.

Decision Tree

The Decision Tree (DT) is a machine learning method that makes decisions based on the graphic structure of a decision tree. In this method, each node of the decision tree represents an attribute, and the tree is created based on the relationship between the attributes. Typically, a decision tree is formed using a set of training data. During training, the Decision Tree algorithm selects the best attribute to split the data based on a metric such as entropy or Gini impurity, which measures the level of impurity or randomness in the subsets. The goal is to find the attribute that maximizes the information gain or the reduction in impurity after the split. By following each ray from the root to the terminal nodes, the samples move from the leaves to the root, and the final classification is determined based on the label of the leaves for each sample. The decision tree method has several advantages, including simplicity and high comprehensibility, the ability to check important features, the ability to use discrete and continuous input data, the ability to estimate any type of feature, and finally, the ability to check and evaluate complex conditions. Decision trees classify the examples by sorting them down the tree from the root to some leaf/terminal node, with the leaf/terminal node providing the classification of the example. The decision tree algorithm can be used for solving regression and classification problems. Decision trees imitate human thinking, so it’s generally easy for data scientists to understand and interpret the results. Decision tree algorithms are powerful tools for classifying data and weighing costs, risks, and potential benefits of ideas [13, 14].

Random Forest

Random forest (RF) is a machine learning method that combines several decision trees. In this method, a random forest consists of several decision trees, each of which is trained independently using a random subset of features and data. The main advantage of random forest is that by combining multi-tree decisions, it avoids single-tree decisions that may be incomplete, innumerable, and highly dependent on the training data. This method can be very useful and powerful for cases where the number of features is large and changeable. The forest generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms. The algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome. The ensemble of trees outputs either the mode or mean of the individual trees, allowing for more accurate and stable results by relying on a multitude of trees rather than a single decision tree [15, 16].

Support vector machine

Support vector machine (SVM) is a machine learning method that has wide applications in the field of computer vision, pattern recognition, classification, and regression. The main working principle of SVM is to separate different data using a decision boundary. In the simplest case, SVM tries to create a decision boundary between the two sets of data. This decision boundary should be defined in such a way that there is the greatest possible distance between the data of each category and the boundary. We call this distance "margin". The performance of SVM in data separation is by training a model using the training data set and finding support vector machines (key points in creating the decision boundary) based on the optimization process. The distance of the closest points from both categories to the border is called the margin, and after training, SVM is able to predict new data using the decision border. One of the interesting features of SVM is that it is capable of semi-supervised or non-linear data separation using a function called the kernel function. Kernel function (such as Gaussian function or polynomial function) maps some data to a higher dimensional space so that linear separation is possible in this space. As a robust and valid algorithm, SVM performs very well in many classification and regression problems. SVM is a powerful machine learning algorithm used for linear or nonlinear classification, regression, and even outlier detection tasks. SVMs can be used for a variety of tasks, such as text classification, image classification, spam detection, handwriting identification, gene expression analysis, face detection, and anomaly detection. SVMs are adaptable and efficient in a variety of applications because they can manage high-dimensional data and nonlinear relationships [17, 18].

Results

Dataset

This dataset includes records related to 171 cases of benign ovarian tumors and 178 records related to cases of ovarian cancer. The age distribution of the two groups is illustrated in Fig. 3. The age of the samples is between 15 and 83, the average age of the samples is 45 years and their standard deviation is 15.1.

Analysis factors affecting the differential diagnosis of ovarian cancer

The effective factors obtained by the Information Gain method in the diagnosis of ovarian cancer malignancy are indicated in Fig. 4 and the Gini Index method in Fig. 5. Three of the most important influential factors in both Information Gain and Gini Index techniques are HE4, CA125, and NEU.

Evaluation of the effectiveness of predictive models for ovarian cancer

A comparison of the performance of these models based on accuracy, sensitivity, specificity, and area under the ROC curve is provided in Table 2. The Random Forest model was identified as the best predictive model with an accuracy of more than 85%. The ROC diagram of the best model (RF) is presented in Fig. 6.

Table 2
Comparing the performance of ovarian cancer prediction models

Model	Accuracy	AUC	F measure	Sensitivity	Specificity
Decision Tree	0.8291	0.799	0.8467	0.8882	0.7636
Random Forest	**0.8675**	**0.925**	**0.8801**	0.9160	**0.8136**
Support Vector Machine	0.8525	0.910	0.8712	**0.9327**	0.7636
AutoMLP (ANN)	0.7935	0.890	0.8169	0.8706	0.7074

Discussion

In this study, the influencing factors on the differential diagnosis of ovarian cancer were investigated, and among the 49 investigated characteristics, the most effective factors were obtained by the Information gain method and Gini index, respectively are HE4, CA125, and NEU, and age.

Also, based on the available data, Decision Tree, Random Forest, Support Vector Machine, and Artificial Neural Network models were generated and compared in terms of accuracy, sensitivity, specificity, f-measure, and AUC parameters, that the random forest model was able to provide the highest performance compared to other models.

Similar studies have been conducted using machine learning methods in the field of ovarian cancer. For example, Jun Ma et al. built models to predict ovarian cancer using machine learning methods based on biomarkers including circulating tumor cells (CTC). The best predictive model of the Random Forest method was reported. This model has been able to predict ovarian cancer with an area under the ROC curve of about 80% [19].

In another study conducted by Lu Pin et al., only the SVM model was used to predict ovarian cancer and its recurrence. The results showed that the group that had a higher response rate to the chemotherapy drug exhibited recurrence in a longer time. The SVM model was able to show a sensitivity of over 90%. The limitation of this study is not using other models and not examining other parameters related to the performance of the model [11].

Also, Haonan Lu et al. developed machine learning models to predict ovarian cancer based on data from medical images of 364 patients. In this study, mathematical descriptors based on machine learning were used. The conclusion was that the descriptors used to predict ovarian cancer did not show a high prognostic power, but due to being non-invasive, and faster than clinical methods, they are of interest [12].

Also, studies such as the study of Akter and his colleagues used methods other than biomarkers based on vaginal ultrasound screening images using machine learning methods have achieved favorable results [20].

The current study could introduce high-performance machine learning models that can be applied to predict ovarian cancer and overcome the limitations and disadvantages of clinical methods.

Conclusion

Various artificial intelligence (AI) tools have emerged as efficient tools based on available data to predict various cancers. These models can help specialists or patients in decision support systems. AI-based algorithms that predict survival and prognosis in cancer patients can be cost-effective and straightforward tools used to support medical decision-making. In this study, it was concluded that different artificial intelligence tools can be used with high accuracy and sensitivity in predicting ovarian cancer. Therefore, the use of these tools can help specialists and patients with early, easier, and less expensive diagnosis of ovarian cancer. AI and machine learning (ML) are increasingly being used in cancer imaging, precision oncology, and cancer diagnosis and screening. AI-based algorithms can predict treatment responses and provide robust computational tools for investigating cancer biology. The application of AI in cancer practice includes providing clinical decision support for cancer diagnosis and screening, processing medical images, and predicting cancer types. AI can help determine where a patient's cancer arose and identify people with the highest risk of pancreatic cancer up to three years before an actual diagnosis. AI can also predict cancer survival rates better than previous tools. In conclusion, AI and machine learning have shown promise in diagnosing, predicting, and potentially even treating a range of medical conditions, including cancer. AI-based algorithms can be cost-effective and straightforward tools used to support medical decision-making.

Declarations

Ethics approval and consent to participate

The project was found to be in accordance to the ethical principles and the national norms and standards for conducting Medical Research in Iran with Approval ID IR.TUMS.SPH.REC.1401.277 evaluated by Research Ethics Committees of School of Public Health & Allied Medical Sciences- Tehran University of Medical Sciences at 27 Feb 2023.

Consent for publication

Not applicable

Availability of data and materials

All data and materials are available. Any data except the data in the manuscript can be provided upon request.

Competing interests

The authors declare that they have no competing interests.

Funding

This study has been funded and supported by Tehran University of Medical Sciences (TUMS); Grant no. 1401-4-102-63824

Authors' contributions

S.M.A and M.A analyzed and interpreted the data, S.M.A performed the analysis of the data, and All authors were contributors in writing the manuscript. All authors read and approved the final manuscript

Acknowledgements

This study has been funded and supported by Tehran University of Medical Sciences (TUMS); Grant no. 1401-4-102-63824

References

1. Huang S, Yang J, Fong S, Zhao Q. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. Cancer letters. 2020;471:61-71.
2. Zugazagoitia J, Guedes C, Ponce S, Ferrer I, Molina-Pinelo S, Paz-Ares L. Current challenges in cancer treatment. Clinical therapeutics. 2016;38(7):1551-66.
3. Wu L, Qu X. Cancer biomarker detection: recent achievements and challenges. Chemical Society Reviews. 2015;44(10):2963-97.
4. Zhu JW, Charkhchi P, Akbari MR. Potential clinical utility of liquid biopsies in ovarian cancer. Molecular Cancer. 2022;21(1):114.
5. Reid BM, Permuth JB, Sellers TA. Epidemiology of ovarian cancer: a review. Cancer biology & medicine. 2017;14(1):9.
6. Jelovac D, Armstrong DK. Recent progress in the diagnosis and treatment of ovarian cancer. CA: a cancer journal for clinicians. 2011;61(3):183-203.
7. Xu H-L, Gong T-T, Liu F-H, Chen H-Y, Xiao Q, Hou Y, et al. Artificial intelligence performance in image-based ovarian cancer identification: A systematic review and meta-analysis. EClinicalMedicine.

2022;53:101662.

8. Stewart C, Ralyea C, Lockwood S. Ovarian cancer: an integrated review. Seminars in oncology nursing: Elsevier; 2019. p. 151-6.
9. Enshaei A, Robson C, Edmondson R. Artificial intelligence systems as prognostic and predictive tools in ovarian cancer. Annals of surgical oncology. 2015;22:3970-5.
10. Shen J, Zhang CJ, Jiang B, Chen J, Song J, Liu Z, et al. Artificial intelligence versus clinicians in disease diagnosis: systematic review. JMIR medical informatics. 2019;7(3):e10010.
11. Lu T-P, Kuo K-T, Chen C-H, Chang M-C, Lin H-P, Hu Y-H, et al. Developing a prognostic gene panel of epithelial ovarian cancer patients by a machine learning model. Cancers. 2019;11(2):270.
12. Lu H, Arshad M, Thornton A, Avesani G, Cunnea P, Curry E, et al. A mathematical-descriptor of tumor-mesoscopic-structure from computed-tomography images annotates prognostic-and molecular-phenotypes of epithelial ovarian cancer. Nature communications. 2019;10(1):764.
13. Gupta B, Rawat A, Jain A, Arora A, Dhami N. Analysis of various decision tree algorithms for classification in data mining. International Journal of Computer Applications. 2017;163(8):15-9.
14. Kotsiantis SB. Decision trees: a recent overview. Artificial Intelligence Review. 2013;39:261-83.
15. Krauss C, Do XA, Huck N. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. European Journal of Operational Research. 2017;259(2):689-702.
16. Hu J, Szymczak S. A review on longitudinal data analysis with random forest. Briefings in Bioinformatics. 2023;24(2):bbad002.
17. Ayyoubzadeh SM, Ghazisaeedi M, Rostam Niakan Kalhori S, Hassaniazad M, Baniasadi T, Maghooli K, et al. A study of factors related to patients' length of stay using data mining techniques in a general hospital in southern Iran. Health information science and systems. 2020;8:1-11.
18. Noble WS. What is a support vector machine? Nature biotechnology. 2006;24(12):1565-7.
19. Ma J, Yang J, Jin Y, Cheng S, Huang S, Zhang N, et al. Artificial intelligence based on blood biomarkers including CTCs predicts outcomes in epithelial ovarian cancer: A prospective study. OncoTargets and therapy. 2021:3267-80.
20. Akter L, Akhter N. Ovarian cancer prediction from ovarian cysts based on TVUS using machine learning algorithms. Proceedings of the International Conference on Big Data, IoT, and Machine Learning: BIM 2021: Springer; 2022. p. 51-61.

Figures

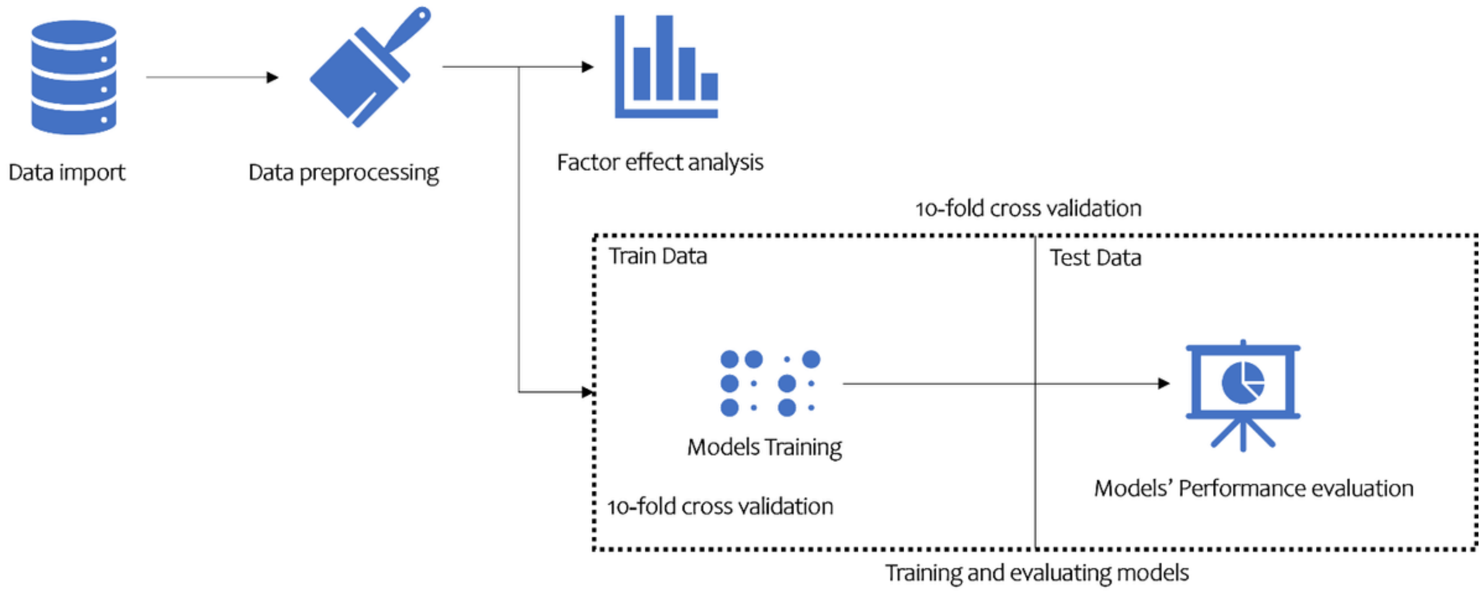

Figure 1

Research Methodology

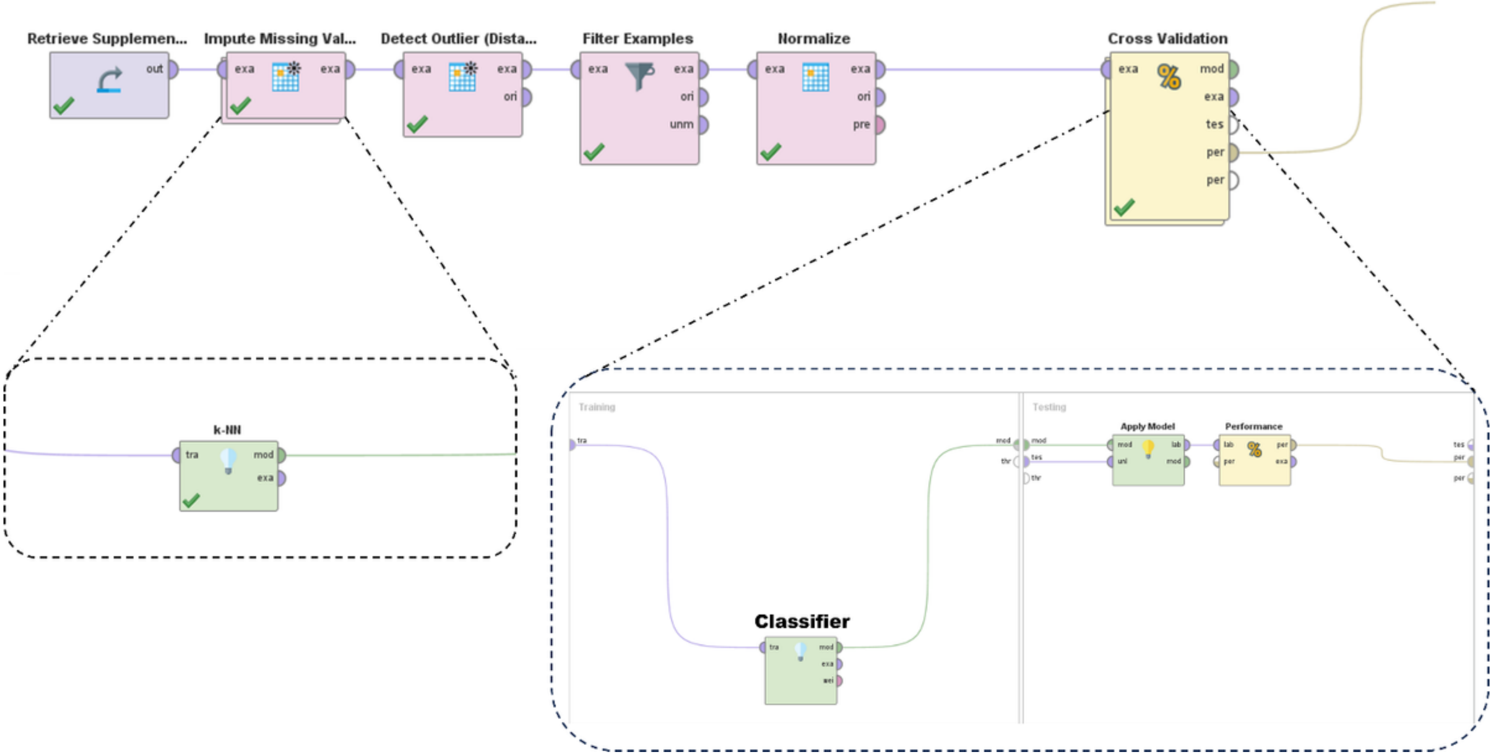

Figure 2

The process designed in RapidMiner software to evaluate predictive models

[redacted]

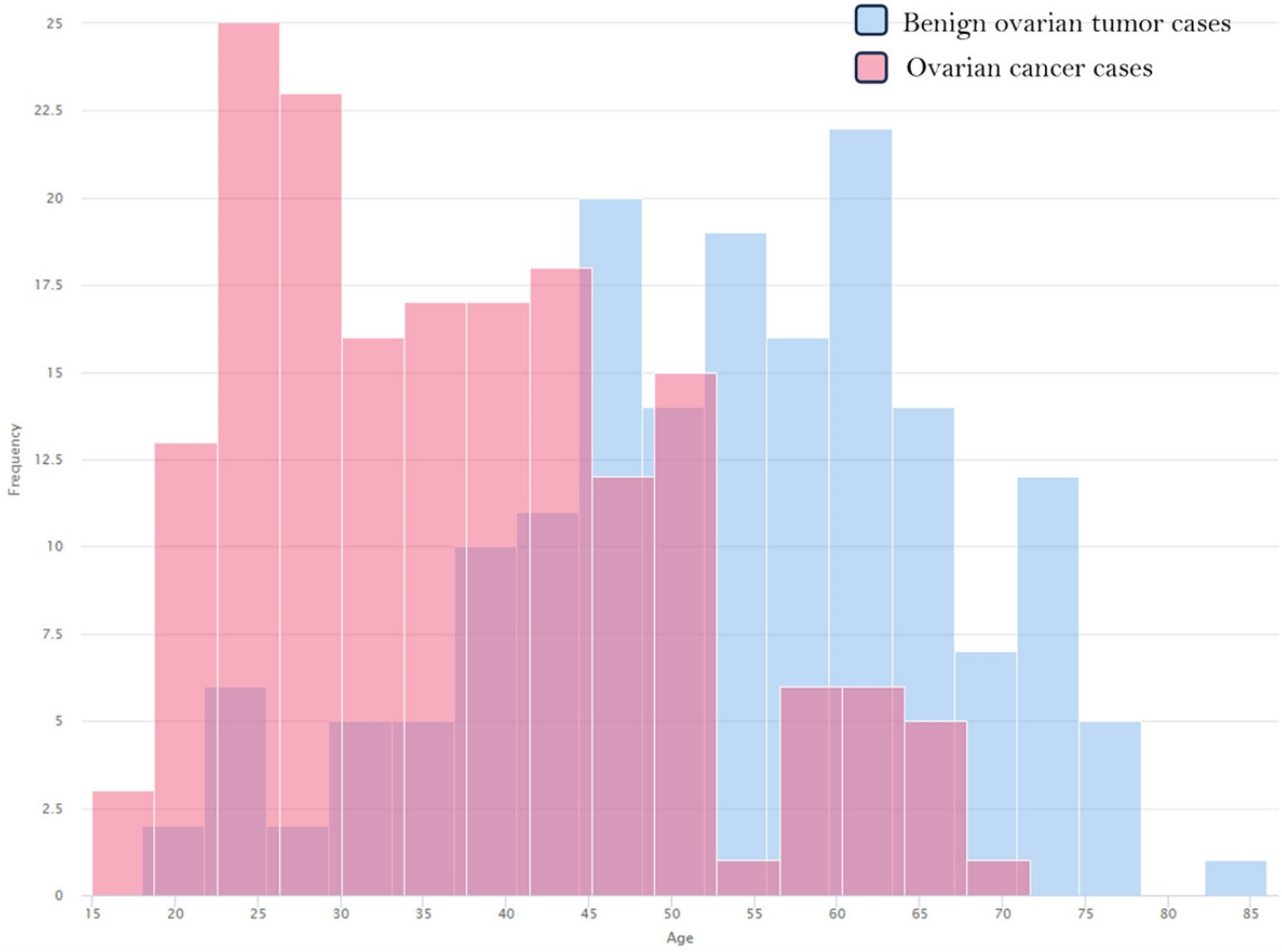

Figure 3

Age distribution of patients with two groups of malignant ovarian cancer and benign ovarian tumor

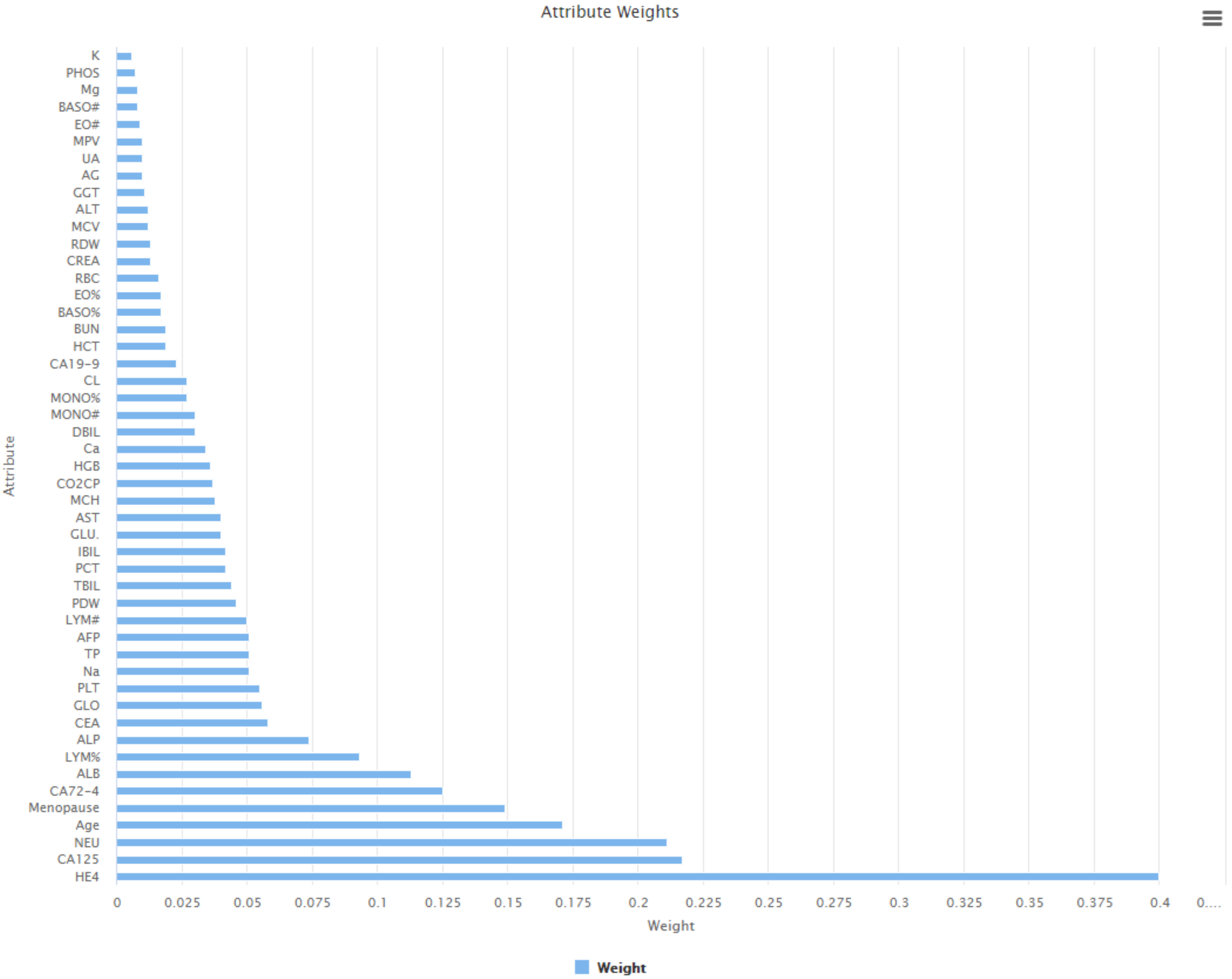

Figure 4

Factors affecting the diagnosis of malignant ovarian cancer obtained by the Information Gain method

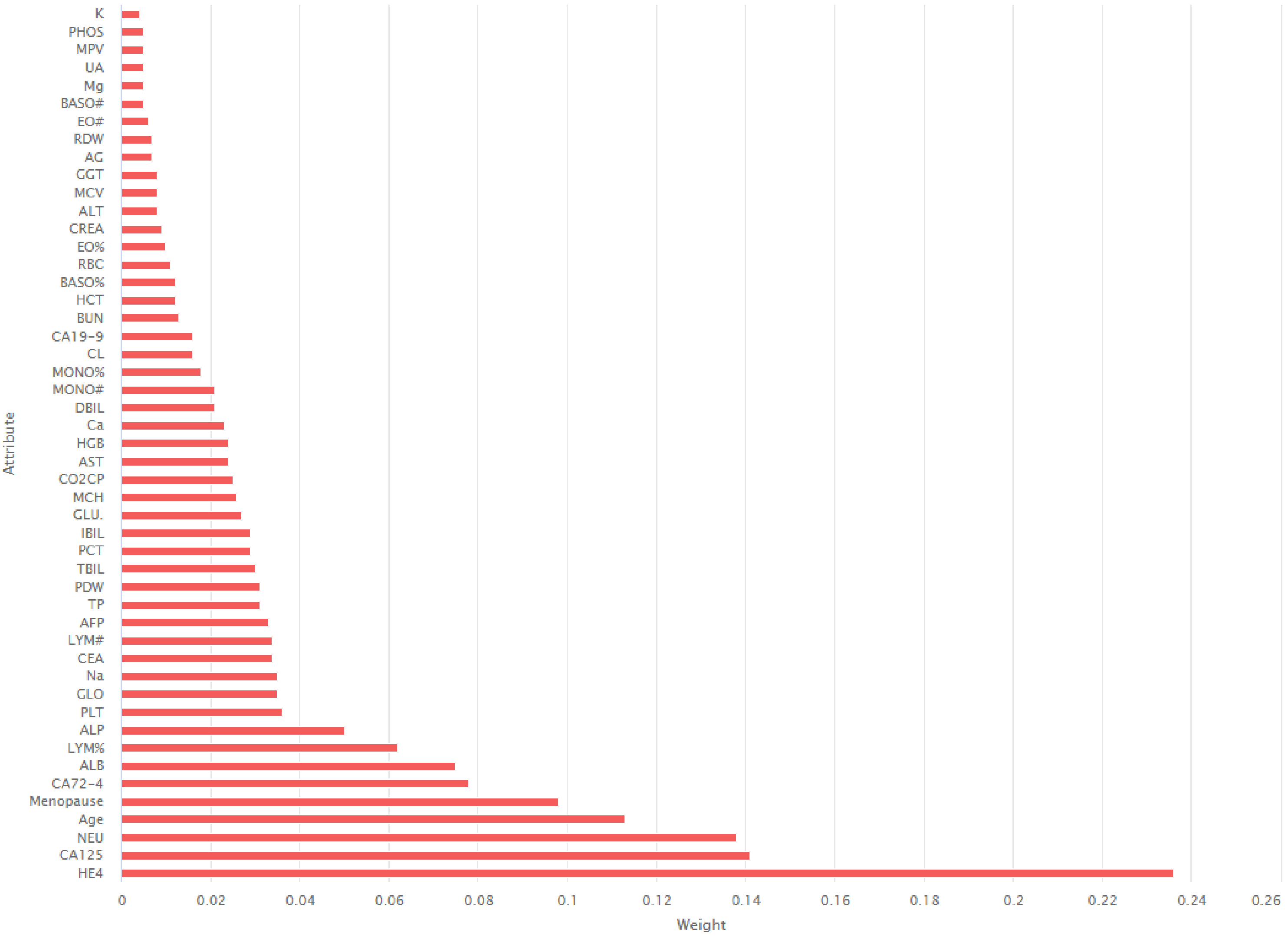

Figure 5

Factors affecting the diagnosis of malignant ovarian cancer obtained by the Gini Index method

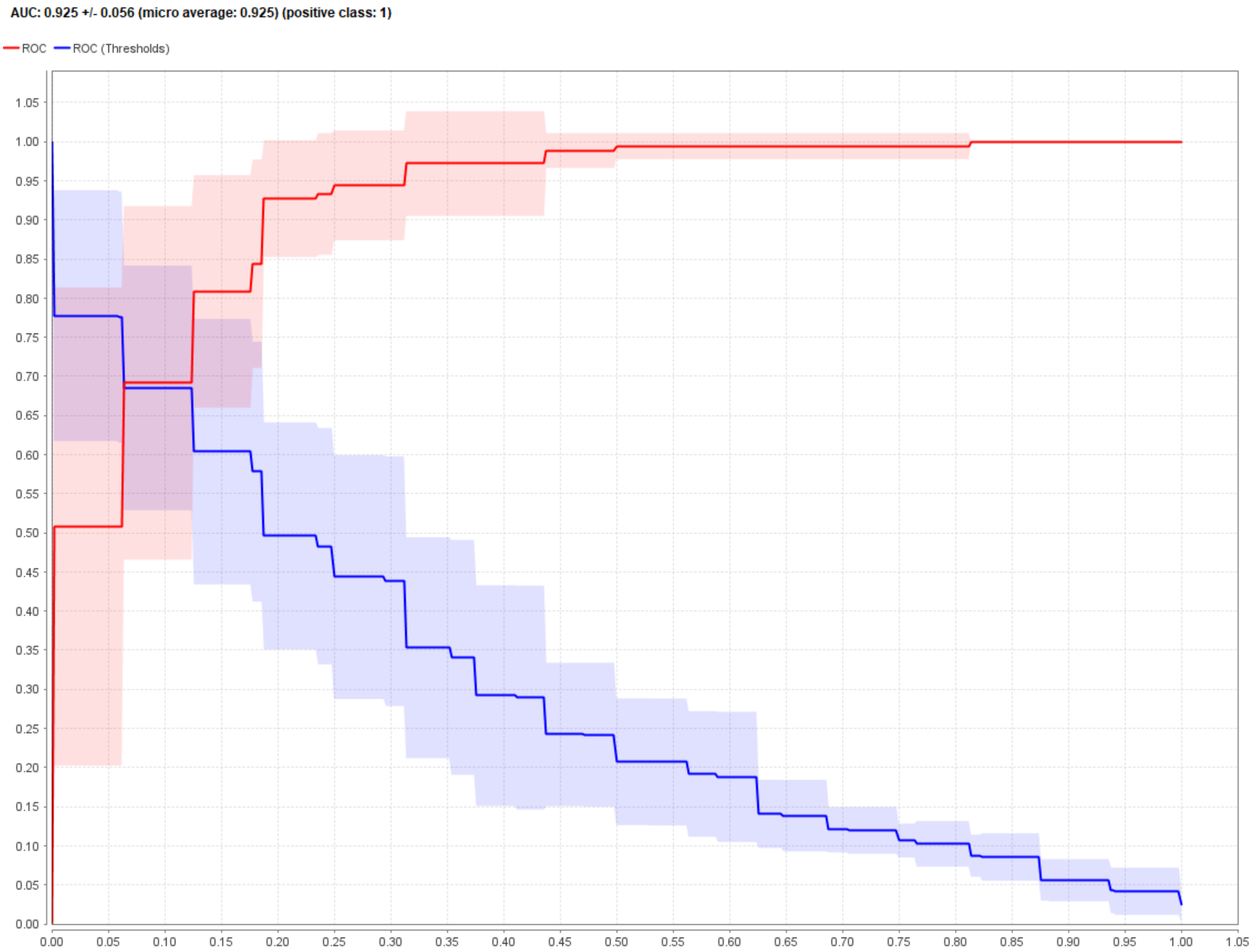

Figure 6

ROC diagram of Random Forest model