# Comparative Study on Different Classification Techniques for Ovarian Cancer Detection

Jasna Nuhić, Lemana Spahić, Sabahudin Ćordić, and Jasmin Kevrić

## Abstract

Diagnosing ovarian cancer is a medical challenge to clinical researchers. This study aims to develop a novel prototype of clinical management in diagnosis and management of patients with ovarian cancer. Various classification algorithms can be applied to cancer databases to devise methods that can predict cancer manifestation. Various methods, however, vary in terms of the level of accuracy, depending on the classification algorithm used. Identifying the most accurate classification algorithm is a challenging task, primarily due to limited data availability. In this paper, a comprehensive comparative analysis of nine different classification algorithms was conducted and their performances have been evaluated. The results indicate that all classifiers are relatively equal in accuracy, meaning that multiple classifying techniques can be used to support physicians in rendering more informed diagnostic decisions.

## Keywords

Ovarian cancer • Classification • Machine learning • Attribute selection

J. Nuhić (✉)
Universitat Pompeu Fabra, Sarajevo,
Bosnia and Herzegovina
e-mail: jasna.nuhic01@estudiant.upf.edu

L. Spahić · S. Ćordić · J. Kevrić
International Burch University, Sarajevo,
Bosnia and Herzegovina
e-mail: lemana.spahic@stu.ibu.edu.ba

S. Ćordić
e-mail: sabahudin.cordic@stu.ibu.edu.ba

J. Kevrić
e-mail: jasmin.kevric@ibu.edu.ba

## 1 Introduction

Cancer is malignancy and is caused by abnormal cell proliferation spreading to other tissue It is not only a name for one particular disease but for a group of related diseases [1]. Cancer is not limited to a single location inside the body but starts as a grouping of mutated cells in a location in the body and spreads throughout it after those malignant cells multiply and enter the bloodstream. Another characteristic is that their cells are less specialized than normal cell due to their inability to mature into a cell type with specific function [2, 3].

Gene expression can be defined as the process by which a gene encoded by DNA is used for synthesis of a functional product that can be either a protein or some other functional product. Control of when and how often genes are expressed plays a crucial role in maintaining homeostasis in the organism [4].

Ovarian cancer accounts for 2.3% of all cancer deaths. In Europe, 65,538 new patients are diagnosed with ovarian cancer each year and 42,716, or 65%, end tragically [5]. In the United States of America, 22,440 women per year are diagnosed with ovarian cancer and 63% of patients die [6].

Diagnosing cancer is a very complex procedure that is susceptible to human and equipment error. First, a biopsy of problematic tissue is conducted and then it is subjected to cytological and molecular tests. These tests are performed in isolated environments in order to minimize potential errors, but errors can always happen [7]. The most effective way to reduce cancer deaths is to detect it earlier.

Rapid advancement of informational technology provided available and, in most cases, inexpensive devices to collect and store the data. In modern, well equipped, clinics data is gathered and shared in large information systems [8]. Nowadays, a vast number of clinical data and patient histories are available for ovarian cancer. With development of biomedical engineering, biomedical engineering researchers

are examining the usage of machine learning techniques in supporting the diagnosis made by medical professionals.

Diagnosis is a relatively straightforward machine learning problem. In this process, a larger set of symptoms and conditions is considered for each patient in respect to classical diagnostic procedure, whereas medical professionals can consider only a limited number of parameters and give a diagnosis based on their interactions [9–11]. By using machine learning and data mining algorithms, medical professionals can establish better diagnoses, choose optimal medications for their patients, predict readmissions, identify patients at high-risk for poor outcomes, and in general improve patient's health while minimizing costs.

The realization of the complexity of certain decisions to treat particular diseases by scientists began 25 years ago when the importance of artificial intelligence became accentuated [12]. Use of data mining techniques will make diagnostic process more agile and also reduce health care costs and waiting times for the patients. The main advantage of using machine learning is the extraction of essential information from a large amount of data and its correlations [13].

Cancer is group of critical areas where classification has a crucial role and where data mining and machine learning are very powerful as tools in medical diagnosis. Therefore, machine learning techniques can help doctors make an accurate diagnosis of ovarian cancer and make the correct classification between cancerous and healthy tissue basing on gene expressions, as well as determining the cancer stage. The most important component in the diagnosis is evaluation of data taken from the patient and the specialist's decision, but artificial intelligence techniques provide support in rendering more informed decisions for medical professionals.

Ovarian cancer diagnosis prediction can be done by integrating big data and machine learning in order to aid prediction, diagnosis and therapy. Gene expression analysis performed by Millstein et al. identified 313 genes that are candidates for involvement in high grade ovarian cancer [14]. A support vector machine was used by Furey et al. to classify cancer tissue samples using gene expression microarrays and they found that most machine learning methods perform comparably on datasets utilized [15]. DNA methylation biomarkers were used as inputs of machine learning by Wei et al. and they have obtained significant results using various classification algorithms [16].

The major contributions of this paper are:

- comparison of different data mining algorithms on ovarian cancer datasets;
- identification of the best performing algorithm to predict ovarian cancer;

- extraction of useful, classified and accurate attributes for prediction of cancer;
- optimization of the task of correctly selecting the set of medical tests that a patient must perform to have the most accurate, the less expensive and time-consuming diagnosis possible;
- propose further investigation for the relevant genes determined in the paper to confirm their role in ovarian cancer detection; and
- obtain new discoveries in ovarian cancer mutation analysis.

This study's aim was to investigate different machine learning techniques. Several algorithms have been used and applied on the ovarian cancer dataset. The focus is on nine machine learning techniques: Naïve Bayes, Multilayer Perceptron, Simple Logistic, Nearest Neighbor, AdaBoost, Attribute Selected Classifier, Random Committee, PART, LMT and Random Forest. These various algorithms were tested using WEKA toolkit and their results analyzed.

## 2 Methods

### 2.1 Dataset

The Ovarian dataset used in this study was obtained from the Gene Expression Omnibus (GEO) database [17]. It consists of 148 samples of female patients. For each patient 83 gene expressions (attributes, features) were collected. Those gene expressions represent attributes, apart from the class attribute (patient having or not having ovarian cancer). Samples were taken invasively by biopsy of tissues of cancer suspicious patients. Gene expressions of those tissues were measured using qPCR, which is the gold standard for gene expression analysis [18, 19]. All diagnosis was performed and confirmed by medical professionals.

The number of samples corresponding to healthy and ovarian cancer groups is presented in Table 1. Of a total of 148 samples, 91 (i.e. 61.5%) had ovarian cancer while the remaining 57 (i.e. 38.5%) did not have ovarian cancer.

### 2.2 Classification Algorithms

Classification, or supervised learning, maps the data into predefined groups and classes. It is performed in two steps: model construction and model usage. Model construction is composed of a set of predetermined classes. The model is constructed by the training set. It is represented as classification rules, decision trees, or mathematical formulas [20].

**Table 1** Data division of ovarian cancer database

| Training dataset | |
| --- | --- |
| Classification group | Number of samples |
| Normal | 57 |
| Cancer | 91 |
| $\sum$ | 148 |

Attributes or unknown objects are classified by comparison of test sample with result from the model. In order to avoid over-fitting, the test set must be independent of the training set. The percentage of correctly classified test samples is known as the accuracy rate [20].

A total of nine classification algorithms were used in this comparative study and the Attribute Selected Classifier for attribute selection. The classifiers have been categorized into different groups such as Bayes, Functions, Lazy, Rules, Tree based classifiers, etc. A mix of algorithms have been chosen from these groups, according to the classification accuracy obtained. In order to obtain better and more robust accuracy, 10-fold cross validation was performed. The following sections briefly explain each of these algorithms.

(a) **Naïve Bayes**

The Naïve Bayes is widely used because of its clarity, elegance, and wholeness, which are reasons for its wide application range. It is combination of Naïve and Bayes, where Naïve stands for independence and Bayes for the Bayes rule. Independence assumes that the attributes are independent of each other [21].

Another assumption is that numeric attributes obey a Gaussian distribution, which is not always true. Therefore, sometimes other methods for estimating continuous distributions are preferred.

(b) **Nearest Neighbor**

Nearest Neighbor is a type of lazy learner classifiers with the main characteristic of storing instances during training. The learning process tends to be slow. The classification itself happens by a majority vote of its neighbours. Nearest Neighbour classifier proved to outperform many other classifiers in two-class problems [22].

(c) **Multilayer Perceptron**

Multilayer Perceptron (MLP) is a class of feed-forward artificial neural network (ANN) with one or more hidden layers between input and output layer. The advantage of such a structure is its ability to avoid overfitting and accomplish nonlinear multiple regressions reliably. MLP's simple architecture can model most nonlinear problems while preserving low computational cost [21].

(d) **Simple Logistic**

Simple Logistic algorithm is a classifier for building linear logistic regression models that also copes quite well with overfitting. Simple logistic algorithms perform much better on dataset with small number of records. However, tree and ensemble tree classifiers can outperform it for larger datasets [21]. Such algorithms are explained further on.

(e) **PART**

PART is a type of rules classifiers and it uses the separate-and-conquer strategy to build a rule. By building a partial decision tree per iteration it does global optimization in order to produce accurate rule sets [22].

(f) **LMT**

LMT is a classifier from the decision trees group, used for building 'logistic model trees' (LMTs). LMTs are classification trees with logistic regression functions at the leaves. The LMT algorithm can deal with binary and multi-class target variables, numeric and nominal attributes and missing values. It ensures that only relevant attributes are included [23].

(g) **AdaBoost**

AdaBoost is a machine learning algorithm that is part of an ensemble methods called boosting where subsequent models attempt to fix the prediction errors made by prior models. It uses short decision tree models, called decision stumps since each has single decision point. The first model is normally constructed, but subsequent models are trained and added until no further improvements are possible [24].

(h) **Random Committee**

Random committee is form of ensemble learning approach. It is based on the assumption of improving performance by combining classifiers. Each classifier construction is denoted

by a different random number of seeds based on the same data. The output class is actually the average of predictions generated by each of these individual base classifiers [22].

### (i)  Random Forest

Random Forest is an ensemble of decision trees that consist of many decision trees. They are form of a nearest neighbor predictor with the output in terms of the mode of the class´s output by individual trees. Random Forest usually yields fast and efficient models due to the possibility of usage without much modeling and handcrafting needed [22, 25].

### (j)  Attribute Selected Classifier for Attribute Selection

When Attribute Selected Classifier is used, the dimensionality of training and test data is reduced by attribute selection before being passed on to a classifier. That ability to select potentially relevant attributes is an essential data engineering component.

Three attribute selection systems used in this study are: locally produced correlation technique, wrapper method and Relief [26]. There are no restrictions for base classifiers.

Correlation based Feature Selection or CFS measures correlation between nominal attributes. It is an automatic algorithm that does not require specification of threshold or number of attributes to be selected. It is assumed that attributes are independent of each other, but strongly related to class. In case that attributes are dependent, there is great possibility for CFS to fail to select all the relevant attributes [25].

The two CFS algorithms used for attribute selection in this study are the CFS Subset Evaluation and the Correlation Attribute Selection. The CFS Subset Evaluation method evaluates the worth of a subset of attributes by considering the individual predictive ability of each attribute, as well as the degree of redundancy between them. The search method used for CFS Subset Evaluation method is Greedy Stepwise. It performs a greedy forward or backward search through the space of attribute subsets. Correlation Attribute Selection evaluation method reduces data by attributes selection before passing it on to a classifier. Search method used for it is Ranker. Ranker ranks attributes by their individual evaluations.

The Wrapper strategy uses an induction algorithm in order to estimate the merit of the attribute. Attribute wrappers are tuned to the specific interaction between an induction algorithm and its data. That makes them perform better than filters, but they tend to be much slower due to the re-run each time different induction algorithm is used [27].

## 3  Results and Discussion

For this study, 37 different classification algorithms were used to diagnose healthy and sick patients. The performance is defined as accuracy, which was determined as: (number of correctly classified samples)/(total number of samples) (Table 2).

For further application, the algorithms with best accuracy were chosen from each group of classifiers. From Bayes classification group, Naïve Bayes classifier was chosen (with an accuracy of 89.25%). From the functions group, Multilayer Perceptron and Simple Logistic classifiers were chosen (with an accuracy of 96.77%). From the Lazy group, Nearest Neighbor classifier was chosen (with an accuracy of 91.3978%). From META group, AdaBoost classifier (with an accuracy of 95.70%) and Random Committee classifier (with an accuracy of 93.55%) were chosen. From RULES group, PART classifier was chosen (with an accuracy of 91.40%). From the Tree group, LMT classifier (with an accuracy of 96.77%) and Random Forest classifier (with an accuracy of 95.67%) were chosen. No algorithm from MISC group is chosen due to its low accuracy. It should be noted that no more than two classifiers from one group were selected.

Table 3 compares and similar studies on ovarian cancer using different databases, but the same or similar algorithm. The performance is calculated by subtracting the result of another study from result obtained in this one. The positive performance indicates that this study outperformed the other, while negative indicate that it underperformed. Out of 13 comparisons, a positive result was obtained in 12 of them and one neutral. Nine of them were outperformed by 5 or more percent, which can be considered a significant difference. The case when it failed to outperform can be attributed to the difference between algorithms compared, since a genetic algorithm was not introduced in this study. Methodologically the most similar one is CV Parameter Selection, which was considered initially in testing phase for this study and later discarded due to its low accuracy.

### Results with Attribute Selection

In order to extract relevant attributes, Attribute Selected Classifier was used. Three different evaluation methods were computed: Correlation-based Feature Selection (CFS) Subset Evaluation, Correlation Attribute Evaluation and Wrapper Subset Evaluation. All previously chosen classifiers were implemented as base classifier for each of evaluation methods. Different search methods were used, and pairs of evaluation and search methods are in Table 4.

**Table 2** Accuracy results for classification algorithms

| Classifying group | Classifying technique | Accuracy (%) |
|---|---|---|
| Bayes | Naïve Bayes | 89.25 |
| Functions | Multilayer perceptron | 96.77 |
| | Simple logistic | 96.77 |
| Lazy | Nearest neighbor | 91.40 |
| Meta | AdaBoost | 95.70 |
| | Random committee | 93.55 |
| Rules | PART | 91.40 |
| Trees | LMT | 96.78 |
| | Random forest | 95.67 |

**Table 3** Studies employing similar machine learning methodologies

| Dataset | Results (x) | Results of this study (y) | y − x (%) |
|---|---|---|---|
| Extrauterine pelvic mass samples (biopsy) [28] | ANN sensitivity → 87.5% | MLP sensitivity → 96.7% | 9.2 |
| | ANN specificity → 92.7% | MLP specificity → 96.8% | 4.2 |
| | Small Naïve Bayes sensitivity → 94.7% | Naïve Bayes sensitivity → 100% | 5.3 |
| | Small Naïve Bayes specificity → 74.1% | Naïve Bayes specificity → 84.1% | 10 |
| | Large Naïve Bayes sensitivity → 96.6% | Naïve Bayes sensitivity → 100% | 3.4 |
| | Large Naïve Bayes specificity → 79.9% | Naïve Bayes specificity → 84.1% | 4.2 |
| Ultrasound description of adnexal pathology [29] | Logistic regression1 sensitivity → 92% | Simple logistic sensitivity → 97% | 5 |
| | Logistic regression1 specificity → 87% | Simple logistic specificity → 97% | 10 |
| | Logistic regression2 sensitivity → 92% | Simple logistic sensitivity → 97% | 5 |
| | Logistic regression1 specificity → 86% | Simple logistic specificity → 97% | 11 |
| Tumor markers and lipid associated sialic acid [30] | Multilayer perceptron accuracy → 90% | Multilayer perceptron accuracy → 97% | 7 |
| Gene expression profiling [31] | k-nearest neighbor accuracy → 71% | IBk accuracy → 91% | 20 |
| Proteomic spectrum data [32] | Genetic algorithm accuracy → 99% | CV parameter selection accuracy → 68% | −31 |

**Table 4** Attribute selection evaluation and search method pairs

| Attribute selection | |
|---|---|
| Evaluation method | Search method |
| CFS subset evaluation | Greedy stepwise |
| Correlation attribute evaluation | Ranker |
| Wrapper subset evaluation | Best first |

In CFS Subset Evaluation and Correlation Attribute Evaluation methods, the same attributes were selected for each base classifier. Selected Classifiers are shown in Table 5. As we can see in Table 4, both methods selected 18 attributes. Out of those 36 selected attributes, genes 22, 23, 24, 30, 58 and 76 (the six of them), were selected by both methods, so there are 30 different relevant attributes according to CFS Subset Evaluation and Correlation Attribute Evaluation methods.

Accuracy of base classifiers applied for the CFS Subset Evaluation and Correlation Attribute Evaluation methods are shown in Table 6. For CFS Subset Evaluation method, the most successful were Nearest Neighbor and Random Forest classifiers with an accuracy of 94.6237%, while for Correlation Attribute Evaluation method AdaBoost was the best performing classifier with an accuracy of 95.6989%.

In the Wrapper Subset Evaluation method, different attributes were selected for each base classifier. Due to the space limitations, the selected classifiers are not presented in tabulated format, while their accuracy can be found in Table 6. The wrapper methods combined with different base classifiers gave different numbers of selected attributes each time performed. There are 30 different attributes selected at least once. Out of those 30 selected attributes genes 23 and 30, namely GYG1p1 and GSK3Bp3, were selected by five different base classifiers, which is the best result.

Four attributes were selected by all three methods, so they are the most relevant attributes for ovarian cancer classification. Those attributes are 22, 23, 24 and 30.

The genes selected by most machine learning techniques were: CALM 1/2/3, GYG1p1, PHKG2p3 and GSK3Bp3. The CALM gene (Calmodulin) encodes for calcium binding proteins that are subunits of phosphorylase kinase, meaning that they are included in cellular signaling [33–35]. GYG1p1 (Glycogenin 1) gene codes for proteins involved in glycosyltransferase that is a catalyst and involved in signaling [36]. PHKG2p3 (Phosphorylase Kinase catalytic subunit Gamma 2) is a gene coding for protein involved in glycogen storage and kinase activity and some variants are overexpressed in cancer [37]. GSK3Bp3 (Glycogen synthase kinase 3 beta) is an enzyme that catalyzes phosphorylation and is found to be upregulated in cancer [38].

## 4  Conclusion

In this study, nine classification techniques, namely Naïve Bayes, Multilayer Perceptron, Simple Logistic, Nearest Neighbor, AdaBoost, Random Committee, PART, LMT, and Random Forest were used to evaluate the percentage of accuracy, with and without attribute selection, for effective prediction techniques for ovarian cancer diagnostics.

**Table 5** Results of attribute selection of CFS subset evaluation and correlation attribute evaluation methods for ovarian cancer database

| Evaluator | Selected attributes | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CF subset evaluator | 10 | 16 | 17 | 22 | 23 | 24 | 30 | 33 | 35 | 36 | 39 | 41 | 58 | 62 | 68 | 76 | 77 | 82 |
| Correlation attribute evaluator | 4 | 3 | 21 | 79 | 22 | 24 | 58 | 5 | 30 | 80 | 23 | 76 | 20 | 25 | 56 | 54 | 26 | 19 |

**Table 6** Attribute selection and simple classification results

| Classifier | | Accuracy (%) | | | |
|---|---|---|---|---|---|
| | | CFS subset evaluation | Correlation attribute evaluation | Wrapper subset evaluation | No attribute selection |
| Naïve Bayes | | 89.25 | 87.10 | 89.25 | 89.25 |
| Multilayer perceptron | | 92.47 | 82.80 | 93.55 | 96.77 |
| Simple logistic | | 88.17 | 86.02 | 87.1 | 96.77 |
| Nearest neighbor | | 94.62 | 93.55 | 97.85 | 91.40 |
| AdaBoost | | 92.47 | 95.70 | 87.10 | 95.70 |
| Random committee | | 92.47 | 90.32 | 89.25 | 93.55 |
| PART | | 92.47 | 90.32 | 92.48 | 91.40 |
| LMT | | 91.40 | 87.10 | 93.55 | 96.78 |
| Random forest | | 94.62 | 93.55 | 91.4 | 95.67 |

In order to achieve the aforementioned objective, an ovarian cancer dataset was utilized. The attribute selection techniques were used to eliminate those attributes that have no significance in the classification process. Therefore, attribute selection technique is the most reliable and the most significant method to improve the accuracy of different classification techniques.

In this study, classification rules were compared to predict the best classifier to develop a new prototype for diagnosis with a predictable pattern for discovery of ovarian cancer. Experimental results show the effectiveness of the proposed method. The base for this is knowledge discovery and data mining. A classifier was identified to determine the nature of the disease, which is highly important for differentiating between healthy and ovarian cancer patients. When compared with previous similar studies using different databases, it clearly shows precedence of the use of GEO database. This study is useful in uncovering patterns hidden in the data that can help the clinicians and doctors in decision making.

All of the mentioned genes are members of signaling cascades and their mutation or disfunction consequently leads to cancer development. Considering the fact that these genes have been determined as relevant by machine learning in this study can lead to further investigation from a biological perspective and experiments that would confirm the prediction. Not only would a positive outcome prove the efficacy of the used ML but could also lead to new discoveries in ovarian cancer mutation analysis.

## References

1. Moses, H.L., Nass, S.J. (eds.): Cancer Biomarkers: The Promises and Challenges of Improving Detection and Treatment. National Academies Press (2007)
2. Kufe, D.W., Pollock, R.E., Weichselbaum, R.R., Bast, R.C., Gansler, T.S.: Holland-Frei Cancer Medicine. BC Decker Inc (2003)
3. Frank, S.A.: Dynamics of Cancer: Incidence, Inheritance, and Evolution. Princeton University Press (2007)
4. Watson, J.D., Baker, T., Bell, S., Gann, A., Levine, M., Losick, R.: Molecular Biology of the Gene. Pearson/Benjamin Cummings (2003)
5. EUCAN Factsheets | Ovarian Cancer. Eco.iarc.fr. Retrieved 10 June 2017, from http://eco.iarc.fr/eucan/CancerOne.aspx?Cancer=27&Gender=2 (2017)
6. What Are the Key Statistics About Ovarian Cancer. Cancer.org. Retrieved 14 June 2017, from https://www.cancer.org/cancer/ovarian-cancer/about/key-statistics.html (2017)
7. Schmidt, U., Begley, C.G.: Cancer diagnosis and microarrays. Int. J. Biochem. Cell Biol. 35(2), 119–124 (2003)
8. Sivagami, P.: Supervised learning approach for breast cancer classification. Int. J. Emerg. Trends Technol. Comput. Sci. 1(4) (2012)
9. Ghaheri, S. Shoar, M. Naderan, Hoseini, S.S.: The applications of genetic algorithms in medicine. Oman Med. J. 30(6), 406 (2015)
10. Fagella.: 7 Applications of Machine Learning in Pharma and Medicine. Available at https://goo.gl/1SlR5k (2017)
11. Magoulas, G.D., Prentza, A.: Machine learning in medical applications. In: Machine Learning and Its Applications. Springer, Berlin Heidelberg, pp. 300–307 (2001)
12. Aloraini, A.: Different machine learning algorithms for breast cancer diagnosis. Int. J. Artif. Intell. Appl. 3(6), 21 (2012)
13. Bin Othman, M.F., Yau, T.M.S.: Comparison of different classification techniques using WEKA for breast cancer. In: 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006, pp. 520–523. Springer, Berlin, Heidelberg (2007)
14. Millstein, J., Budden, T., Anglesio, M., Talhouk, A., Beeghly-Fadiel, A., Berchuck, A., García, M.: A gene expression prognostic signature for overall survival in patients with high-grade serous ovarian cancer (2018)
15. Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D.: Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16(10), 906–914 (2000)
16. Wei, S.H., Balch, C., Paik, H.H., Kim, Y.S., Baldwin, R.L., Liyanarachchi, S., Karlan, B.Y.: Prognostic DNA methylation biomarkers in ovarian cancer. Clin. Cancer Res. 12(9), 2788–2794 (2006)
17. Shen, L., Zeng, Q., Guo, P., Huang, J., Li, C., Pan, T., Huang, T.: Dynamically prognosticating patients with hepatocellular carcinoma through survival paths mapping based on time-series data. Nat. Commun. 9(1), 2230 (2018)
18. VanGuilder, H.D., Vrana, K.E., Freeman, W.M.: Twenty-five years of quantitative PCR for gene expression analysis. Biotechniques 44(5), 619–626 (2008)
19. Derveaux, S., Vandesompele, J., Hellemans, J.: How to do successful gene expression analysis using real-time PCR. Methods 50(4), 227–230 (2010)
20. Aher, S.B., Lobo, L.M.R.J. Data mining in educational system using weka. In: International Conference on Emerging Technology Trends (ICETT), vol. 3, pp. 20–25 (Mar 2011)
21. Kumar, Y., Sahoo, G.: Analysis of Bayes, neural network and tree classifier of classification technique in data mining using WEKA (2012)
22. Chetty, G., White, M., Akther, F.: Smart-phone-based data mining for human activity recognition. Procedia Comput. Sci. 46, 1181–1187 (2015)
23. Landwehr, N., Hall, M., Frank, E.: Logistic model trees. Mach. Learn. 59(1–2), 161–205 (2005)
24. Freund, Y., Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In Icml (Vol. 96, pp. 148–156)
25. Nookala, G.K.M., Pottumuthu, B.K., Orsu, N., Mudunuri, S.B.: Performance analysis and evaluation of different data mining algorithms used for cancer classification. Int. J. Adv. Res. Artif. Intell. 2(5), 49–55 (2013)
26. Witten, I.H., Frank, E., Trigg, L.E., Hall, M.A., Holmes, G., Cunningham, S.J.: Weka: practical machine learning tools and techniques with Java implementations (1999)
27. Kohavi, R., GH, John: Wrappers for feature subset selection. Artif. Intell. 97(1–2), 273–324 (1997)
28. Antal, P., Verrelst, H., Timmerman, D., Moreau, Y., Van Huffel, S., De Moor, B., Vergote, I.: Bayesian networks in ovarian cancer diagnosis: potentials and limitations. In: Computer-Based Medical Systems, CBMS 2000. Proceedings of 13th IEEE Symposium, pp. 103–108. IEEE (2000)
29. Kaijser, J., Bourne, T., Valentin, L., Sayasneh, A., Van Holsbeke, C., Vergote, I., Timmerman, D.: Improving strategies for diagnosing ovarian cancer: a summary of the International Ovarian Tumor Analysis (IOTA) studies. Ultrasound Obstet. Gynecol. 41(1), 9–20 (2013)

30. Zhang, Z., Zhang, H., Bast, R.C.: An application of artificial neural networks in ovarian cancer early detection. In: Neural Networks, IJCNN 2000. Proceedings of the IEEE-INNS-ENNS International Joint Conference, vol. 4, pp. 107–112. IEEE (2000)

31. Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., Yakhini, Z.: Tissue classification with gene expression profiles. J. Comput. Biol. **7**(3–4), 559–583 (2000)

32. Stevens, E.V., Liotta, L.A., Kohn, E.C.: Proteomic analysis for early detection of ovarian cancer: a realistic approach? Int. J. Gynecol. Cancer **13**, 133–139 (2003)

33. CALM1 calmodulin 1 [Homo sapiens (human)]. Retrieved from https://www.ncbi.nlm.nih.gov/gene/801 (2018)

34. CALM2 calmodulin 2 [Homo sapiens (human)]. Retrieved from https://www.ncbi.nlm.nih.gov/gene/805

35. CALM3 calmodulin 3 [Homo sapiens (human)]. Retrieved from https://www.ncbi.nlm.nih.gov/gene/808 (2018)

36. GSK3B glycogen synthase kinase 3 beta [Homo sapiens (human)]. Retrieved from https://www.ncbi.nlm.nih.gov/gene/2932 (2018)

37. PHKG2 phosphorylase kinase catalytic subunit gamma 2 [Homo sapiens (human). Retrieved from https://www.ncbi.nlm.nih.gov/gene/5261 (2018)

38. GYG1 glycogenin 1 [Homo sapiens (human)]. Retrieved from https://www.ncbi.nlm.nih.gov/gene/2992 (2018)