ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2019.0278

RISK PREDICTION SYSTEM USING DATA MINING TECHNIQUES IN GYNECOLOGICAL OVARIAN CANCER

Vidyaa Thulasiraman[1] and S. Kavitha[2]
[1]*Department of Computer Science, Government Arts and Science College for Women, Bargur, India*
[2]*Department of Computer Science, Auxilium College, India*

Abstract

Cancer is one of the leading causes of death worldwide. Early detection and prevention of cancer plays a very important role in reducing deaths caused by cancer. Ovarian Cancer (OC) is a type of cancer that affects ovaries in women, and is difficult to detect at initial stage due to which it remains as one of the leading causes of cancer death. Identification of genetic and environmental factors is very important in developing novel methods to detect and prevent cancer. This research uses data mining technology such as classification, clustering and prediction to identify potential cancer patients. Therefore a cancer risk prediction system is here proposed which is easy, cost effective and time saving.

Keywords:

Ovarian Cancer, Multi-Layer Perceptron Classifier, Detection

1. INTRODUCTION

Ovarian cancer is the leading cause of death from Gynecological malignancies with an estimated 65,697 new cases and 41,448 deaths each year in Europe [1]. Approximately 15% of women present with disease localized to the ovaries and in this group with full staging surgery the 5-year survival is >90%. However, the majority of women present with advanced disease (International Federation of Gynecological Oncology (FIGO) stage III-IV) and their survival at 5 years is poor, currently <30%. Early diagnosis is fundamental to achieving a high cure rate, but this is difficult due to the paucity of clearly defined symptoms. At present, there is no evidence for screening asymptomatic women although trials are in progress. Advanced ovarian cancer is most commonly diagnosed following presentation with symptoms and some of these may be present in early-stage disease.

Most women with early ovarian cancer are cured by surgery. Ovarian cancer is contained epithelial ovarian, essential peritoneal and fallopian tube carcinoma [1] [2]. After initial treatment, most patients with ovarian cancer have undetectable diseases and are thought to be in clinical abatement. Cancer is a potentially fatal disease caused mainly by environmental factors that mutate genes encoding critical cell-regulatory proteins. The resultant aberrant cell behavior leads to expansive masses of abnormal cells that destroy surrounding normal tissue and can spread to vital organs resulting in disseminated disease, commonly a harbinger of imminent patient death.

More significantly, globalization of unhealthy lifestyles, particularly cigarette smoking and the adoption of many features of the modern Western diet (high fat, low fiber content) will increase cancer incidence. [3] Detecting cancer is still challenging for the doctors in the field of medicine. Even now the actual reason and complete cure of cancer is not invented. Various tests are available for predicting cancer, but detecting cancer in earlier stage is difficult, but earlier detection of cancer is curable. We have proposed the cancer prediction system based on data mining. Cancer prediction system estimates the risk of the gynecologic cancer especially in ovary.

Ovarian cancer is cancer that begins in the ovaries. Ovaries are reproductive glands establish only in women. The ovaries produce eggs (ova) for reproduction. The egg's journey during the Fallopian tubes into the uterus where the fertilized egg embeds and establishes into a fetus. The ovaries are also the major cause of the female hormones estrogen and progesterone. One ovary is situated on each side of the uterus in the pelvis. Many types of tumors can generate rising in the ovaries.

The majority of these are benign (noncancerous) and never multiply outside the ovary. Benign tumors can be treated effectively by removing either the ovary or the part of the ovary that contains the tumor. Ovarian tumors that are not benign or malignant (cancerous) and can increase (metastasize) to other parts of the body. Ovarian tumors are named according to the kind of cells the tumor in progress from and whether the tumor is benign or cancerous. There are 3 main types of ovarian tumors: Epithelial tumors establish from the cells that wrap the outer surface of the ovary. Most ovarian tumors are epithelial cell tumors. Germ cell tumors begin from the cells that generate the eggs (ova). Stromal tumors begin from structural tissue cells that grip the ovary collectively and make the female hormones estrogen and progesterone.

A widely recognized formal definition of data mining can be defined as "Data mining is the non-trivial extraction of implicit previously unknown and potentially useful information about data". Data mining has some fields to analysis of data such as classification, clustering, correlations, association rule etc. [4] and has been used intensively and extensively by many organizations.

Data mining technique involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods in early detection of cancer. In classification learning, the learning scheme is presented with a set of classified examples from which it is expected to learn a way of classifying unseen examples.

In association learning, any association among features is sought, not just ones that predict a particular class value. In clustering, groups of examples that belong together are sought [5]. In numeric prediction, the outcome to be predicted is not a discrete class but a numeric quantity. Data Mining techniques are implemented together to create a novel method to diagnose the existence of cancer for a particular patient. When beginning to work on a data mining problem, it is first necessary to bring all the data together into a set of instances. Integrating data from different sources usually presents many challenges.

2. LITERATURE SURVEY

Jeetha and Malathi [6] intended to observe performance of Artificial Neural Network (ANN) over genetic algorithm on diagnosis of ovarian cancer. The conclusion of this study was to use ANN along with genetic algorithm and propose a method for refinement and categorizing the ovarian cancer with kind, spreading, and normal tissue. Since studies have focused on how a genetic algorithm performs when compared to machine learning technique, it is interesting to learn how these techniques be used as an advantage to learn about survival of ovarian cancer patients using non-modifiable factors.

Cervical cancer is a disease that affects 2,66,000 deaths worldwide and is the fourth highest incidence of cancer in women. This cancer can be diagnosed through a Pap smear, where a cytopathologist observes a microscopic image of the cervix cells to determine whether the patient is normal or abnormal. The sensitivity and specificity of the Pap smear is known to be respectively 53.4% and 69.2%. Since the test is related to the patient's life, it is important to improve the accuracy of the test [7].

A variety of systems have been proposed to help judge experts to improve the accuracy of tests in the medical field, but the development of these systems has been limited to areas where digitized test data are clearly present. In this paper, we design and propose a model that automatically classifies normal/abnormal states of cervical cells from microscopic images using convolutional neural network and several machine learning classifiers. As a result, the support vector machine showed the best performance with a 78% F1 score [7].

In this study, ensemble learning and five data mining approaches, including support vector machine (SVM), C5.0, extreme learning machine (ELM), multivariate adaptive regression splines (MARS), and random forest (RF) are integrated to rank the importance of risk factors and diagnose the recurrence of ovarian cancer. The medical records and pathologic status were extracted from the Chung Shan Medical University Hospital Tumor Registry. Experimental results illustrated that the integrated C5.0 model is a superior approach in predicting the recurrence of ovarian cancer. Moreover, the classification accuracies of C5.0, ELM, MARS, RF, and SVM indeed increased after using the selected important risk factors as predictors. Our findings suggest that The International Federation of Gynecology and Obstetrics (FIGO), Pathologic M, Age, and Pathologic T were the four most critical risk factors for ovarian cancer recurrence. In summary, the above information can support the important influence of personality and clinical symptom representations on all phases of guide interventions, with the complexities of multiple symptoms associated with ovarian cancer in all phases of the recurrent trajectory [8].

The objective of this study is to determine if a commercially available classification algorithm biomarker patterns software (BPS), which is based on a classification and regression tree (CART), would be effective in discriminating ovarian cancer from benign diseases and healthy controls. Serum protein mass spectrum profiles from 139 patients with ovarian cancer, benign pelvic diseases, or healthy women were analyzed using the BPS software. A decision tree using five protein peaks resulted in an accuracy of 81.5% in the cross-validation analysis and 80% in a blinded set of samples in differentiating the ovarian cancer from the control groups. The potential, advantages, and drawbacks of the BPS system as a bioinformatics tool for the analysis of the SELDI high-dimensional proteomic data are discussed [9].

In this paper proposed a novel approach for identifying ovarian cancer using combined Self Organizing Maps Immune Clonal Selection (SOMICS) and Grammatical Evolution Neural Networks (GENN). SOMICS algorithm used for better feature selection which is used for extracting valuable, implicit, and interesting information from vast amount of medical data and GENN is used for classification process. The experimental results show the comparison of the proposed method and other classification methods using three various classifiers such as, Support Vector Machine (SVM), Multi Layer Perceptron (MLP), Feed Forward Neural Network (FFNN). The combined SOMICS and GENN method yields promising results on classification and feature selection accuracy for ovarian cancer dataset with classification accuracy of 98.23% mean square error of 0.0021% [10].

This research implements decision tree classifiers and artificial neural network to predict whether the patient will live with ovary cancer or not. Dataset was obtained from Danish Cancer Register and contains five Input parameters. Dataset contains some missing values and a noticeable improvement in accuracy was detected after removing them. Three features of the original dataset were shown to be the most significant: Mobility of the cancer, Surface of the cancer, and the Consistency of the cancer. The addition of the other two features (Size of the cancer and age of the patient) did not improve the results significantly. It was noticed that the patients with a cystic, but fixed and even cancer have always died from the ovary cancer. In contrast, the patients with uneven, but fixed and solid cancer have always survived the cancer. It is recommended to include more information about either the cancer or the patient to increase the chance of predicting the output of such patients [11].

Early detection is paramount to reduce the high death rate of ovarian cancer. Unfortunately, current detection tool is not sensitive. New techniques such as deoxyribonucleic acid (DNA) micro-array and proteomics data are difficult to analyze due to high dimensionality, whereas conventional methods such as blood test are neither sensitive nor specific. Thus, a functional model of human pattern recognition known as Complementary Learning Fuzzy Neural Network (CLFNN) is proposed to aid existing diagnosis methods. In contrast to conventional computational intelligence methods, CLFNN exploits the lateral inhibition between positive and negative samples. Moreover, it is equipped with autonomous rule generation facility [12].

An example named Fuzzy Adaptive Learning Control Network with Another Adaptive Resonance Theory (FALCON-AART) is used to illustrate the performance of CLFNN. The confluence of CLFNN-microarray, CLFNN-blood test and CLFNN-proteomics demonstrate good sensitivity and specificity in the experiments. The diagnosis decision is accurate and consistent. CLFNN also outperforms most of the conventional methods. This research work demonstrates that the confluence of CLFNN-DNA micro-array, CLFNN-blood tests, and CLFNN-proteomic test improves the diagnosis accuracy with higher consistency. CLFNN exhibits good performance in ovarian cancer diagnosis in general. Thus, CLFNN is a promising tool for clinical decision support [12].

Data mining is an essential step in the process of knowledge discovery in databases in which intelligent methods are applied in order to extract patterns. Cancer research is generally clinical and/or biological in nature. Data driven statistical research has become a common complement. Predicting the outcome of a disease is one of the most interesting and challenging tasks with data mining applications (the use of computers powered with automated tools). Large volumes of medical data are being collected and made available to the medical research groups. As a result, Knowledge Discovery in Databases (KDD), which includes data mining techniques, has become a popular research tool for medical researchers to identify and exploit patterns and relationships among large number of variables, and made enable them predict the outcome of a disease using the historical cases stored within datasets. It gives an overview of the current research being carried out on various cancer datasets using the data mining techniques to enhance cancer diagnosis and prognosis [13].

Novel methods, both molecular and statistical, are urgently needed to take advantage of recent advances in biotechnology and the human genome project for disease diagnosis and prognosis. Mass spectrometry (MS) holds great promise for biomarker identification and genome-wide protein profiling. It has been demonstrated in the literature that biomarkers can be identified to distinguish normal individuals from cancer patients using MS data. Such progress is especially exciting for the detection of early-stage ovarian cancer patients. Although various statistical methods have been utilized to identify biomarkers from MS data, there has been no systematic comparison among these approaches in their relative ability to analyze MS data [14].

Compare the performance of several classes of statistical methods for the classification of cancer based on MS spectra. These methods include: linear discriminant analysis, quadratic discriminant analysis, K-Nearest Neighbor classifier, Bagging and Boosting Classification Trees, Support Vector Machine and random forest (RF). The methods are applied to ovarian cancer and control serum samples from the National Ovarian Cancer Early Detection Program clinic at Northwestern University Hospital. We found that RF outperforms other methods in the analysis of MS data [16].

3. RESEARCH METHODOLOGY

The existing method of the SVM is mostly utilized in arrangement of problem, and it managed machine learning algorithm, and it is utilized for classification and regression challenges. In the algorithm, each data item is a point in n-dimensional space (where *n* is the number of features) with the estimation of every component being the estimation of a particular sort out. Support Vectors are only the co-ordinates of individual discernment. Support Vector Machine is a wilderness which best disconnects the two classes (hyper-plane or line).

The proposed method to diagnosis of ovarian cancer disease using intelligent systems attracted the attention of researchers. The proposed systems implemented using techniques called MLP. It based on several databases of patients to obtain a high precision diagnosis. In this work, an architecture data mining technique based cancer prediction system combining the prediction system with mining technology was used. In this model we have used one of the classification algorithms called Multi-layer perceptron (MLP). The best prediction model has been identified named as Multi-Layer Perceptron (MLP) which gives a high accuracy.

Classification is a technique that predicts categorical class labels. It classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data. Therefore a cancer risk prediction system is here proposed which is easy, cost effective and time saving. Current limitations of biomarkers for ovarian cancer screening relate to the relatively poor sensitivity and specificity for detection of early stage disease.

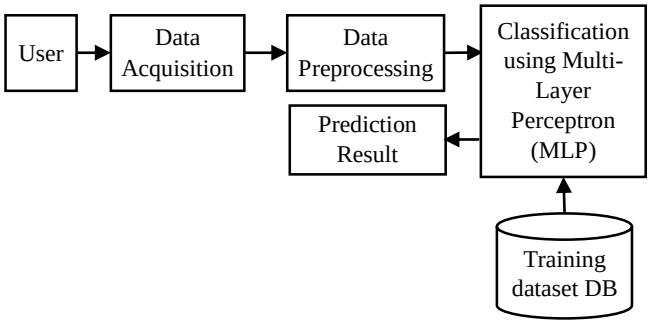

Fig.1. Proposed Architecture

3.1 DATA PREPROCESSING

Data preprocessing is a vital task of data mining. It mainly used for making analysis appropriate and also making data appropriate for clustering by avoiding duplicate records and adding missing data according to past recorded data. The main benefits of data pre-processing reduces memory. The pre-processing stage removes the unwanted data also handle the missing data's in the subset.

3.2 CLASSIFICATION

Predicts the class of objects whose class label is unidentified. Its purpose is to find an obtained model that describes and distinguishes data classes or concepts. The Derived Model is based on the study set of training data i.e. the data object whose class label is well known.

3.3 PREDICTION

It is used to predict missing or engaged numerical data values rather than class labels. Regression study is usually used for prediction. Prediction can also be used for recognition of distribution trends based on available data.

3.4 MULTILAYER PERCEPTRON (MLP)

MLP is one of the most frequently used neural network architectures in Medical Decision Support System (MDSS), and it belongs to the class of supervised neural networks. A typical MLP network consists of three or more layers of processing nodes: an input layer that receives external inputs, one or more hidden layers, and an output layer which produces the classification results (Fig.2).

The fundamental of the neural network is that when data are accessible at the input layer, the network neurons carry out calculations in the hidden layers until an output value is obtained at each of the output neurons. This output indication should be

able to point out the suitable class for the input data. That is, one can expect to have a high output value on the right class neuron and low output values on all the rest. A node in MLP can be modeled as an artificial neuron as shown in Fig.3 that computes the weighted sum of the inputs at the presence of the bias and passes this sum through the activation function.

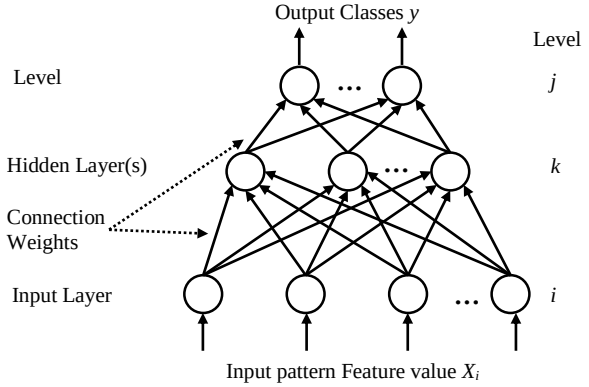

Fig.2. Structure of a Multilayer Perceptron Network

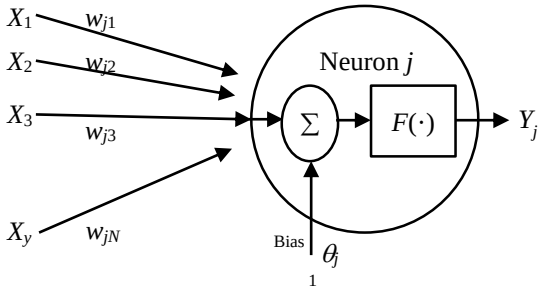

Fig.3. One node of MLP - An Artificial Neuron

The whole process is defined as follows:

$$V_j = \sum_{i=1}^{p} W_{ji} X_i + \theta_i \quad (1)$$

$$y_i = f_j(V_j) \quad (2)$$

where,

V_j is the linear combination of inputs X_1, X_2.

X_p is the bias: The connection weight between the input and the neuron j

$f_i(\cdot)$ is the activation function of the j^{th} neuron

y_i is the output.

The sigmoid function is a common choice of the activation function as defined as.

$$F(\alpha) = \frac{1}{1+e^{-\alpha}} \quad (3)$$

Any network consists of training phase and testing phase. The reason of training phase is to update the weights value of the network based on the learning algorithm. The weights of the neural network are updated in a supervised mode using the most common algorithm known as the Back Propagation (BP) algorithm and according to Eq.(3).

$$W_{ij}(t+1) = W_{ji}(t) - \varepsilon \frac{\delta E_f}{\delta W_{ji}} \quad (4)$$

where, ε is the learning rate and E_f is the error function.

In the learning phase, the output of the feed forward neural network is computed for each input training pattern. The error between the computed output and desired output is used to update the weight of the network by back propagation algorithm.

The working of multilayer perceptron neural network is summarized in steps as mentioned below:

Step 1: Input data is provided to input layer for processing, which produces a predicted output.

Step 2: The predicted output is subtracted from actual output and error value is calculated.

Step 3: The network then uses a Backpropagation algorithm which adjusts the weights.

Step 4: For weights adjusting it starts from weights between output layer nodes and last hidden layer nodes and works backwards through network.

Step 5: When back propagation is finished, the forwarding process starts again.

Step 6: The process is repeated until the error between predicted and actual output is minimized.

4. PERFORMANCE ANALYSIS

The ovarian cancer is the most serious cancer for the woman which leads to death at the saviour stage. This cancer is classified by using Multi-layer perceptron (MLP). In this paper, we present a widespread indication of numerous proposed cancer classifications method and estimate them based on their calculated time and recognition accuracy.

Table.1. Description of Parameters

Attribute	Description
Age	Age in years
Menu pause	Mensural cycle
thestbps	Resting blood pressure
fbs	Fasting blood sugar
ch	cholesterol
es	estrogen
Bleeding_type	Bleeding Type normal and irregular
Bt	Bloating
PAP test	PAP smear at age frequency <21yrs, >21yrs, 21-29, 30-65, above 65
Bl_test	Blood Test

4.1 DATABASE

- Total Patient = 216
- Ovarian Cancer Patient = 121
- Normal patient = 95

- In training 113 patients samples used. (In that 63 samples are cancer and 50 sample normal)
- In testing 103 patients samples used. (In that 58 samples are cancer and 45 sample normal)

4.1.1 Accuracy:

Accuracy is the major factor in cancer classification, where it alone only the objective that wants to achieve.

$$\text{Accuracy} = \left(\frac{\text{No. of sample correctly classified}}{\text{Classified sample result}} \right)$$

4.1.2 Sensitivity:

Specificity is calculated by number of positive samples correctly predicted by total number of positive samples.

$$\text{Sensitivity} = \left(\frac{\text{No. of positive samples correctly predicted}}{\text{Total No. of positive samples}} \right)$$

4.1.3 Specificity:

Specificity is calculated by number of negative samples correctly predicted by total number of negative samples.

$$\text{Specificity} = \left(\frac{\text{No. of negative samples correctly predicted}}{\text{Total No. of negative samples}} \right)$$

4.1.4 Training Time:

Training time is calculated time period taken for algorithm to perform in the system. A training set must contain a list of objects with known classifications.

4.1.5 Precision and Recall:

Precision is the fraction of relevant instances among the retrieved instances, while recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. Precision and Recall are used as a measurement of the relevance.

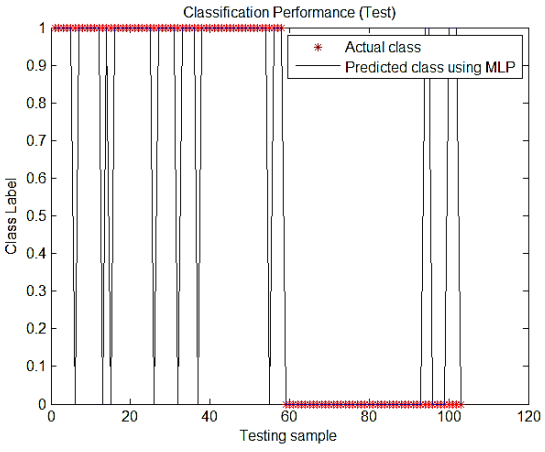

Fig.4. Classification using MLP

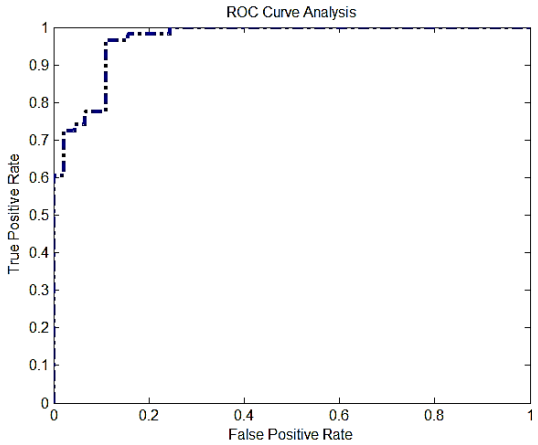

Fig.5. ROC Curve analysis using *TP* and *FP* rate

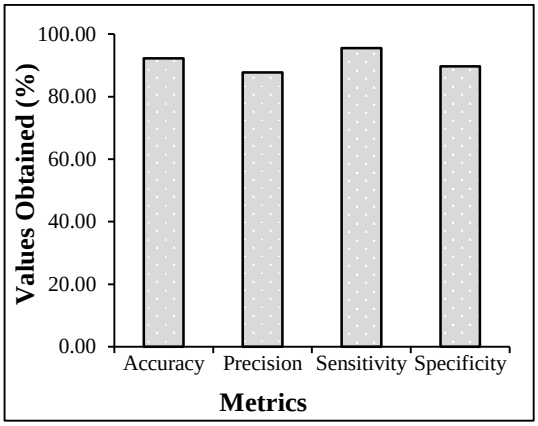

Fig.6. Comparison of various parameters between the proposed system and existing system

Table.2. Classification of Normal and Abnormal Class

Actual Class	Classified Class	
	Normal	Ovarian Cancer
Normal	43	2
Ovarian Cancer	6	52

Table.3. Comparison of various parameters between the proposed system and existing system

Classification Performance	Existing Method (SVM)	Our Method (MLP)
Accuracy	89.32%	92.23%
Precision	85.41%	87.76%
Sensitivity (or) Recall	91.11%	95.56%
Specificity	87.93%	89.66%
Execution Time	0.761451s	0.361371s

5. CONCLUSIONS

The diagnosis system is proposed in this work to assist physicians to diagnosis the condition by converting medical factors of the patients into numerical representation. The

performance results show that, the proposed MLP classifier has 98% accuracy of ovarian diseases classification when the performance of this classifier was evaluated using collected database. Ovarian cancer is one of the critical cancers for the women which lead to death in the critical stage. The pre-processing stage removes the unwanted data also handle the missing data's in the subset. Finally the classification is done with the help of the grammatical evaluation neural network with highest classification accuracy. Thus, the paper proposed that new algorithm to identify and classify the cancer then the performance of the system is evaluated with the help of the experimental result.

REFERENCES

[1] B.T. Hennessy, R.L. Coleman and M. Markman, "Ovarian Cancer", *Lancet*, Vol. 374, No. 9698, pp. 1371-1382, 2009.

[2] J.O. Schorge, S.C. Modesitt, R.L. Coleman, D.E. Cohn, N.D. Kauff, L.R. Duska and T.J. Herzog, "SGO White Paper on Ovarian Cancer: Etiology, Screening and Surveillance", *Gynecologic Oncology*, Vol. 119, No. 1, pp. 7-17, 2010.

[3] S. Jothi and S. Anitha, "Data mining Classification Techniques Applied of Cancer Disease-A Case Study using Xlminer", *International Journal of Engineering Research and Technology*, Vol. 1, No. 8, pp. 23-32, 2012.

[4] T. Jayalakshmi and A. Santhakumaran, "A Novel Classification Method for Classification of Diabetes Mellitus using Artificial Neural Networks", *Proceedings of International Conference on Data Storage and Data Engineering*, pp. 159-163, 2010.

[5] P. Ramachandran, N. Girija and T. Bhuvaneswari, "Early Detection and Prevention of Cancer using Data Mining Techniques", *International Journal of Computer Applications*, Vol. 97, No. 13, pp. 1-8, 2014.

[6] B. Rosiline Jeetha and M. Malathi, "Diagnosis of Ovarian Cancer using Artificial Neural Network", *International Journal of Computer Trends and Technology*, Vol. 4, No. 10, pp. 3601-3606, 2013.

[7] J. Hyeon, H.J. Choi, B.D. Lee and K.N. Lee, "Diagnosing Cervical Cell Images using Pre-Trained Convolutional Neural Network as Feature Extractor", *Proceedings of International Conference on Big Data and Smart Computing*, pp. 411-417, 2017.

[8] C.J. Tseng, C.J. Lu, C.C. Chang and G.D. Chen, "Integration of Data Mining Classification Techniques and Ensemble Learning to Identify Risk Factors and Diagnose Ovarian Cancer Recurrence", *Artificial Intelligence in Medicine*, Vol. 78, pp. 47-54, 2017.

[9] Antonia Vlahou, John O. Schorge, Betsy W. Gregory and Robert L. Coleman, "Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data", *Journal of Biomedicine and Biotechnology*, Vol. 5, No. 1, pp. 308-314, 2003.

[10] P. Yasodha and N.R. Ananthanarayanan, "Detecting the Ovarian Cancer using Big Data Analysis with Effective Model", *Biomedical Research*, Vol. 31, No. 1, pp. 1-14, 2018.

[11] Ahmed Osmanoviu, Layla Abdel Ilah, Adnan Hodziu, Jasmin Kevric and Adnan Fojnica, "Ovary Cancer Detection using Decision Tree Classifiers based on Historical Data of Ovary Cancer Patients", *Proceedings of the International Conference on Medical and Biological Engineering*, pp. 344-349, 2017.

[12] Tuan Zea Tan, Chai Quek, Geok See Ng and Khalil Razvi, "Ovarian Cancer Diagnosis with Complementary Learning Fuzzy Neural Network", *Artificial Intelligence in Medicine*, Vol. 43, pp. 207-222, 2008.

[13] Pooja Agrawal, Suresh Kashyap, Vikas Chandra Pandey and Suraj Prasad Keshri, "Knowledge Patterns in Clinical Data through Data Mining: A Review on Cancer Disease Prediction", *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, Vol. 2, No. 4, pp. 56-64, 2013.

[14] Baolin Wu, Tom Abbott, David Fishman, Walter McMurray, Gil Mor, Kathryn Stone, David Ward, Kenneth Williams and Hongyu Zhao, "Comparison of Statistical Methods for Classification of Ovarian Cancer using Mass Spectrometry Data", *Bioinformatics*, Vol. 19, No. 13 pp. 1636-1643, 2003.

[15] H. Yan, Y. Jiang, J. Zheng, C. Peng and Q. Li, "A Multilayer Perceptron-based Medical Decision Support System for Heart Disease Diagnosis", *Expert Systems with Applications*, Vol. 30, No. 2, pp. 272-281, 2006.

[16] M. Azarbad, S. Hakimi and A. Ebrahimzadeh, "Automatic Recognition of Digital Communication Signal", *International Journal of Energy, Information and Communications*, Vol. 3, No. 4, pp. 21-34, 2012.