Automatic Diagnosis of Ovarian Cancer Based on Relative Entropy and Neural Network

Zainab Abbas Harbi [1]
[1]Department of Computer Science, Colledge of Education for Women, Kufa University, Iraq
zainab.aljuburi@uokufa.edu.iq

Abstract— Ovarian Cancer is one of the most common causes of death for women in developing countries. Screening and early diagnoses of OC are urgently needed. Early diagnosis would help in consequence procedures and treatment. Mass spectrometry (MS) data is been used as an effective component of cancer diagnosis tools. However, these valuable data have a large number of dimensions that can affect the learning process in addition to time-consuming considerations. Feature selection plays an important role in reducing information redundancy, and deals with the invalidation that occurs in basic classification algorithms when there are too many features and huge datasets. To improve the automatic system diagnosis accuracy, entropy-based selection features are proposed. These features are combined with the novel learning capabilities of neural networks to achieve higher diagnostic accuracy. In order to show the performance of the proposed system, experiments have been performed using different feature selection algorithms and machine learning classification approaches. The final results show that the proposed system had 97.7% accuracy and performs better than other approaches.

Keywords— Relative entropy, neural network, ovarian cancer, and automatic diagnosis.

1. Introduction

Ovarian Cancer is the eighth most frequent cause of death due to cancer in women worldwide [1]. 21410 new cases and 13770 deaths have been reported in 2021 [2]. The majority of patients (58%) had advanced diagnoses. Ovarian cancer is a particularly deadly condition because of the typical late stage upon diagnosis. Early diagnosis is the best course of action to raise the low survival rate of ovarian cancer [3].

To create sensitive screening tests that would enable earlier detection of OC, numerous research studies and clinical trials have been carried out [4]. Using a novel technology called Mass Spectrometry (MS), proteomics researchers can quickly and precisely analyze a huge number of proteins in cells and tissues to uncover specific biomarkers associated with cancer. Using MS techniques in conjunction with bioinformatics tools, pathological investigations and disease treatment utilizing protein biomarkers will be improved [5].

Machine learning approaches are useful for identifying common patterns in data, which aids in improving assessments, predictions, and decision-making in a variety of medical cancer diagnosis domains [6]. Many well-known classification Machine Learning (ML) models, including Support Vector Machine (SVM), Linear

Discriminant Analysis (LDA), and Random Forest (RF), have already been applied to and evaluated in cancer-related diagnosis [7,8]. These ML techniques are used on preprocessed MS data; however, the preprocessing variations present a significant obstacle to any comparison analysis and could result in the dimensionality curse. When processing MS data with a lot of dimensions, the well-known dimensionality problem occurs. Techniques for reducing the dimensionality of the data are used to remove unnecessary or redundant features [9]. Several researchers have investigated the diagnostic use of mass spectra utilizing various feature extraction techniques depending on genetic algorithms, data mining, wavelet transforms, and principle components in [10,11,12,13] respectively. The classification performance reported in these studies ranges from 82% to 98%. However, selecting the best features still raises areas of concern. The best feature subset from each feature set is chosen through the process of feature selection. Three algorithms—filter, wrapper, and embedded—have been developed in recent years to be used in the research of feature selection area[14].

It is very important to choose just those features that are necessary for classification to make the most use of the data and minimize the dimension of the data set, and rank the features in both data sets—Cancer and Normal. We have employed a Kullback-Leibler Divergence [15], commonly known as Relative Entropy for feature selection. Then, we used the top n genes for the task of classification, where n is the total number of features included in the classification model. In various classification tasks, neural network models have been employed successfully in the diagnosis of cancer [16]. For the detection and treatment of cancer, classification is essential. Using the selected features with the powerful learning capabilities of the neural network, the automatic diagnosis of ovarian cancer disease can be improved which lead to better therapy consequence and procedure outcomes. In this study, we used publically accessible MS clinical data from the National Cancer Institute FDI-NCI center database [17], and we attempted to categorize the data into groups of cancer and healthy individuals using entropy-based selected features and neural networks.

The remainder of this paper is organized as follows: In section 2, the material and methodologies are presented. The proposed system results and discussions are detailed in section 3. Finally, section 4 concludes the paper and presents the future direction

2. Material and methodologies

The proposed model of ovarian cancer automatic diagnosis consists of four steps as illustrated in Figure 1. The first step is dataset preprocessing, then features selection, classifier modeling and the last step is model evaluations.

2.1. Data Set and Preprocessing

The clinical data set used in this paper was provided by the FDA-NCI center. The National Ovarian Cancer Early Detection Program (NOCEDP) at Northwestern University (Chicago, IL, USA) in the FDA-NCI center, provided the serum samples [17]. The clinical data collection included 216 samples, including 95 samples from healthy individuals and 121 samples from ovarian cancer patients.

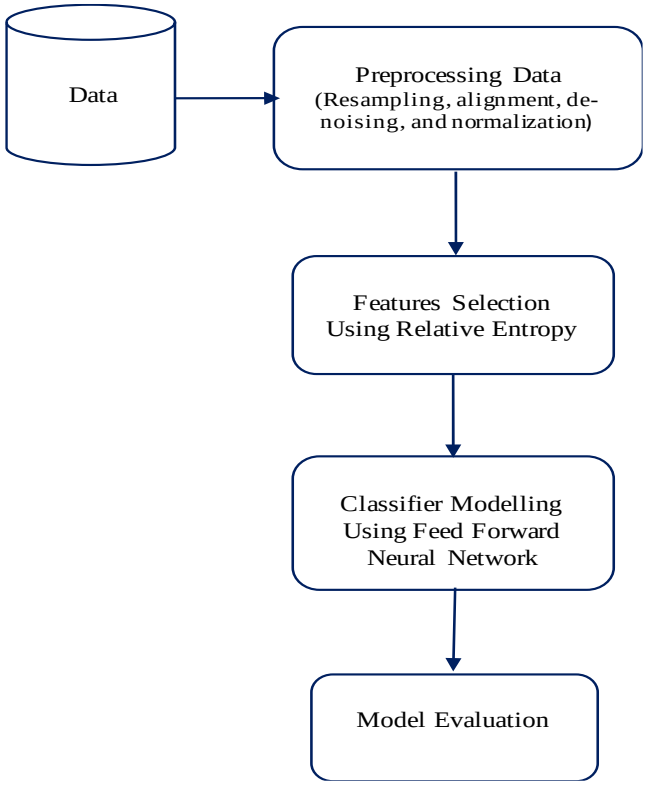

Figure 1. Flow diagram of the proposed mode

The raw data set sample's features have a large dimension. There are around 360,000 attributes per sample. The raw spectrum, however, includes a lot of noise and redundancy, and the differences between the healthy sample and the cancer sample are focused on a small area. These datasets were often subjected to a preprocessing method that included resampling, alignment, de-noising, and normalization. The reader can find a thorough explanation of the preparation technique in [18]. The significant peaks are aligned, the backdrop is adjusted, the dimension is decreased to 15000, and the noise is reduced. In this research, the normalized and preprocessed data set is used.

2.2. Features Selection

The goal of feature selection techniques is to choose the features that contain the majority of the target variable's information. To prevent information redundancy in the supplied set, the chosen features should be independent. Relative Entropy, another name for the Kullback-Leibler divergence, is a filter-based feature selection technique where the features are chosen independently of any machine learning algorithm. The

highest discriminating information and variance are identified, and features are ranked accordingly [19]. Entropy measures the average uncertainty of a random variable *x*. The entropy of a variable *x* is given by:

$$H(x) = -\sum_{x \in X} p(x)\ log_2\ p(x) \quad (1)$$

where *p(x)* is the probability that *X* is in the state *x*. The fundamental issue with the entropy approach is that it ignores the amount of data required to discriminate between a sample and a population. The distribution of variance between and within classes is required. The Kullback-Leibler distance entropy, is an interesting metric that quantifies the differences between two probability distributions . Let *p(x)* and *q(x)* be the probability functions for the discrete distributions *P* and *Q*, respectively. Then the Kullback-Leibler Distance or relative entropy of p to q, is defined by:

$$kLD(p \| q) = \sum_{x \in X} p(x)\ log_2\ \frac{p(x)}{q(x)} \quad (2)$$

where *p(x*) and *q(x)* are two probability mass functions.

In this study, we present a classification approach based on relative entropy. We take into account information from the normal patient and the cancer data sets and assess their relative entropy. The reduced subset of features that are the most important ones is formed by choosing the top n features (n determined by the user's selection). Next, a neural network is used to classify data employing this subset of features.

2.3. Classifier Modeling

A classifier receives the selected features from the features selection step. Neural networks are a machine learning approach that is based on the way neurons communicate with one another in the human brain. Neural networks are often employed to perform pattern recognition and classify items in a variety of disciplines, including images, voice, vision, and control systems [20]. They are particularly well suited for modeling non-linear relationships.

The Multi-Layer Perceptron is the most fundamental type of neural network (Figure 2). Despite having a very simple structure MLP is still a useful model for the majority of classification concerns [21]. MLP is a feed-forward neural network, that has an input layer, one or more hidden layers, and an output layer. Several nodes that are coupled by weights make up each layer. The algorithm modifies the weights of the network throughout the model training phase to increase classification accuracy.

The model training involves several forward and backward passes. Data is transferred from the input layer to the output layer through the network during the forward pass. Partial derivatives of the cost function to the weights are generated by the algorithm in the backward pass and those results are used to modify the weight values.

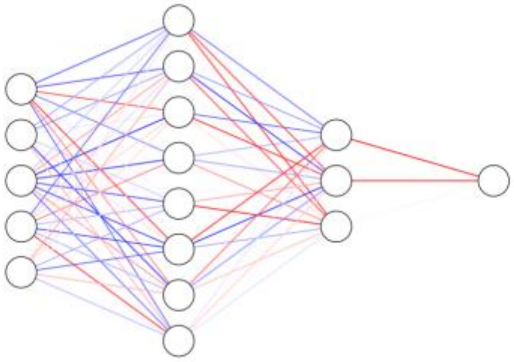

Figure 2. The Architecture of Multi-Layer Perceptron

A single forward pass during the training phase involves calculating the node values for subsequent layers beginning with the input layer. The number of input features is represented in the number of nodes in the input layer. The nodes of the first hidden layer receive the input features. The weighted sum of the input values in addition to a bias term is then transformed using a nonlinear activation function ReLU, as follows:

$$f(x) = \begin{cases} 0 & for\ x < 0 \\ x & for\ x \geq 0 \end{cases} \quad (3)$$

where *x* is the input to a neuron. The ReLU is the most used activation function in the current research. It is used in almost all convolutional neural networks or deep learning. This process is continued until output layer node(s) are computed. A classification task's cost function is calculated based on the mutual information between the predicted and actual values of the target variables. When it comes to classification, the number of output layer neurons is equal to the number of classifying's categories. In our case is two-class cancer and normal.

In addition to neural networks, several classification models are employed in this paper such as LDA, Decision tree, Naïve Bayes, SVM, and KNN. All the models used here are well-known classifiers in machine learning [22].

2.4 Model Evaluation

The study's evaluation phase is a crucial stage. In the evaluation phase, we assess how well the learning models are doing. The learning models can be assessed using a variety of evaluation parameters. The common evaluation factors for cancer detection are utilized in this research and they are well-known and often used in similar areas [23]. Which are recall, precision, and accuracy.

The confusion matrix provides the values that are used to determine the classifier's performance on the test data. False Positive (*FP*), True Negative (*TN*), True Positive (*TP*), and False Negative (*FN*) are used to calculate the evaluation parameters. *Fp* and *TP* stand for false and true positive classification. *FN* and *TN* stand for false and true negative classification. A well-known and often used metric for assessing classifier performance is *Accuracy*, and it is calculated using:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Precision and *Recall* are also other parameters that are commonly used for the classifier performance evaluation. Precision and recall take into consideration the positive cases and can be calculated as:

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

3. Results and discussion

This study compares the effectiveness of the proposed approach using several experiments. All experiments are performed on an Intel Corei7 7th generation computer with Windows 10. Matlab 2021 is used to implement the proposed model. Experiments are conducted independently. In order to verify the model efficiency 5-fold cross-validation is used. The neural network architecture consists of one fully connected layer with 25 neurons and a ReLU activation function

Firstly, the experiments are conducted with a similar number of features (200 features) and a neural network as a classifier. These features are selected with different approaches including the proposed one from the ovarian cancer dataset. The reported results are shown in Table 1. It is clear that the higher accuracy was reported by entropy-based selected features with a classification accuracy of 97.7% in comparison with other approaches. Ttest[24] achieved 95.3% accuracy, while 93% classification accuracy was recorded using Bhattacharyya Distance also known as Chernoff Bound[25].

The neural networks achieved higher accuracy with only 200 Enropy-based selected features out of 1500 original features. The selected features improve the classification performance in terms of time constraints as only 19 sec are needed to train the neural network with selected features in comparison with the 4628 sec required to train the network with original features. The result was reported with ideal Precision of 1.00 and a very high recall of 0.95. The confusion matrix for the test classification result is shown in Figure 3.

Table 1. Classification accuracy results using different features selection techniques

Selected features techniques	Accuracy
Ttest	95.3%
Chernoff bound	93%
Relative Entropy	97.7%

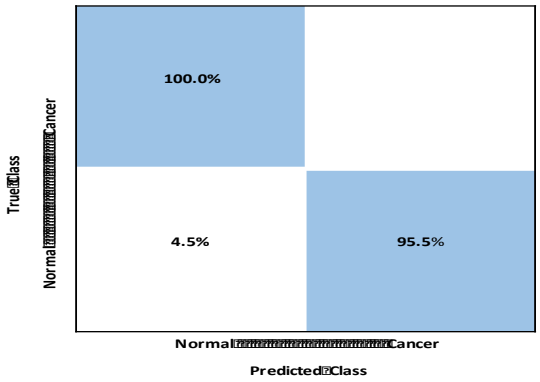

Figure 3. Confusion matrix (Precision and Recall)

Secondly, the experiments are conducted with different classifiers to show the performance of the selected neural network in combination with the selected entropy-based features. Table 2 shows the performance analysis of different classifiers in terms of classification accuracies. As reported in Table 2, The Neural network achieved higher accuracy 97.7% with a similar number of selected entropy-based features compared with another classifier. Both SVM and KNN also achieved good accuracy of 95.3%, while 90.7% and 86% accuracies are attained by decision tree and Naïve Bays classifier respectively. However, a low accuracy value of 74.4 % is achieved in LDA Classifier with the same set of selected features.

Table 2. Classification accuracy results using different classification techniques

Classifier Model	**Accuracy**
LDA	74.4%
Naïve Bayes	86%
Decision tree	90.7%
SVM	95.3%
KNN	95.3%

Neural Network	**97.7%**

To compare the performance of the proposed approach with the existing models that investigate ovarian cancer detection research area, a performance comparison is carried out with the existing studies that are applied to the same data set. For this purpose, several recent studies from the literature are selected and the comparison results are shown in Table 3. For example, in [13] Authors use probabilistic Principle Component Analysis PCA feature's selection with an SVM classifier for ovarian cancer detection and show a 90.8% accuracy. The proposed model in this paper demonstrates better results in comparison with [13] and more than 7% improvement are achieved. In [26] 95% accuracy was reported by transforming data using a wavelet transform combined with a feed forward neural network as a classifier. A convolutional neural network with transfer learning is used in [27] to obtain a 98% accuracy. However, this reported result represents the best result over 10 iterations as stated by the authors, and in our case, we take the average. On the other hand, we achieved comparable results with a small difference of 0.3% in comparison with the complex architecture of convolutional neural networks.

Table 3. Performance comparison with other approaches based on classification accuracy

Reference	**Accuracy**
[13]	90.8%
[26]	95%
[27]	98%
Proposed	97.7%

4. CONCLUSIONS

In this paper, an ovarian cancer automatic diagnosis system is proposed. A set of entropy-based selected features have been used to train a neural network classifier for successfully classifying normal and cancer patients. Comparative experiments show significant improvements in classification accuracy using the proposed feature selection method. Applying the proposed model to a mass spectrum ovarian cancer data set

reported better performance with 97.7 % diagnosis accuracy and an optimal precision of 1.00 and a very high recall of 0.95. These significant results are due to the discriminant features that are extracted using the proposed feature selection technique. In addition, we have shown the effectiveness of neural networks in classification with comparing to other classification approaches such as SVM, KNN, Naïve Bays, Decision Tree, LDA.

The proposed system's usefulness is demonstrated by its better accuracy when compared to alternative methods. However, data pre-processing is not considered in this paper. In future works, better performance could be achieved by exploring preprocessing and data analysis. In addition we aim to discover the effectiveness of features that are selected via a deep-learning approach.

5. References

[1] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 71(3), 209-249.

[2] Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. (2022). Cancer statistics, 2022. CA: a cancer journal for clinicians, 72(1), 7-33.

[3] Menon, U., Gentry-Maharaj, A., Burnell, M., Singh, N., Ryan, A., Karpinskyj, C., & Parmar, M. (2021). Ovarian cancer population screening and mortality after long-term follow-up in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. The Lancet, 397(10290), 2182-2193.

[4] Badgwell D, Bast RC Jr. Early detection of ovarian cancer. Dis Markers. 2007;23(5-6):397-410. doi: 10.1155/2007/309382. PMID: 18057523; PMCID: PMC3851959.

[5] Hossain, K. R., Escobar Bermeo, J. D., Warton, K., & Valenzuela, S. M. (2022). New approaches and biomarker candidates for the early detection of ovarian cancer. Frontiers in Bioengineering and Biotechnology, 10, 157.

[6] Sebastian, A. M., & Peter, D. (2022). Artificial Intelligence in Cancer Research: Trends, Challenges and Future Directions. Life, 12(12), 1991.

[7] Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Zhao, H. (2003). Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics, 19(13), 1636-1643.

[8] Vervier, K., Mahé, P., Veyrieras, J. B., & Vert, J. P. (2015). Benchmark of structured machine learning methods for microbial identification from mass-spectrometry data. arXiv preprint arXiv:1506.07251.

[9] Hilario, M., & Kalousis, A. (2008). Approaches to dimensionality reduction in proteomic biomarker studies. Briefings in bioinformatics, 9(2), 102-118.

[10] Emanuel F Petricoin, Ali M Ardekani, Peter J Levine Ben A Hitt, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, and Lance A Liotta. Use of proteomic patterns in serum to identify ovarian cancer. Lancet, 359(9306):572– 577, 2002.

[11] Lihua Li, Hong Tang, Zuobao Wu, Jianli Gong, Michael Gruidl, Jun Zou, Melvyn Tockman, and Robert A Clark. Data mining techniques for cancer detection using serum proteomic profiling. Artif Intell Med, 32(2):71–83, 2004.

[12] Marina Vannucci, Naijun Sha, and Philip J. Brown. Nir and mass spectra classification: Bayesian methods for wavelet-based feature selection. Chemometrics and Intelligent Laboratory Systems, 77(1):139–148, 2005

[13] Wu, J., Ji, Y., Zhao, L., Ji, M., Ye, Z., & Li, S. (2016). A mass spectrometric analysis method based on ppca and svm for early detection of ovarian cancer. Computational and mathematical methods in medicine, 2016.

[14] Jović, A., Brkić, K., & Bogunović, N. (2015, May). A review of feature selection methods with applications. In 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO) (pp. 1200-1205). Ieee.

[15] Cover, T. M., & Thomas, J. A. (1991). Entropy, relative entropy and mutual information. Elements of information theory, 2(1), 12-13.

[16] Agrawal, S., & Agrawal, J. (2015). Neural network techniques for cancer prediction: A survey. Procedia Computer Science, 60, 769-774.

[17] https://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp, Clinical Proteomics Data Bank, 2014

[18] Conrads, T. P., Fusaro, V. A., Ross, S., Johann, D., Rajapakse, V., Hitt, B. A.,& Veenstra, T. D. (2004). High-resolution serum proteomic features for ovarian cancer detection. Endocrine-related cancer, 11(2), 163-178.

[19] Bommert, Andrea, et al. "Benchmark for filter methods for feature selection in high-dimensional classification data." Computational Statistics & Data Analysis 143 (2020): 106839.

[20] Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. Heliyon, 4(11), e00938.

[21] Desai, M., & Shah, M. (2021). An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN). Clinical eHealth, 4, 1-11.

[22] Kotsiantis, Sotiris B., Ioannis D. Zaharakis, and Panayiotis E. Pintelas. "Machine learning: a review of classification and combining techniques." Artificial Intelligence Review 26.3 (2006): 159-190.

[23] Munir, K., Elahi, H., Ayub, A., Frezza, F., & Rizzi, A. (2019). Cancer diagnosis using deep learning: a bibliographic review. Cancers, 11(9), 1235.

[24] Montgomery, D. C., Runger, G. C., & Hubele, N. F. (2009). Engineering statistics. John Wiley & Sons.
[25] Choi, E., & Lee, C. (2003). Feature extraction based on the Bhattacharyya distance. Pattern Recognition, 36(8), 1703-1709.
[26] Seddik, A. F., Hassan, R. A., & Fakhreldein, M. A. (2013). Spectral Domain Features for Ovarian Cancer Data Analysis. J. Comput. Sci., 9(8), 1061-1068.
[27] Seddiki, K., Saudemont, P., Precioso, F., Ogrinc, N., Wisztorski, M., Salzet, M., & Droit, A. (2020). Cumulative learning enables convolutional neural network representations for small mass spectrometry data classification. Nature communications, 11(1), 5595.

Article submitted 2 Jun 2023. Accepted at 4 July 2023

Published at 30 September 2023.