# *A Novel Spectral Feature Selection Method Based on Binary Genetic Algorithm for Efficient Detection of Endometrial and Ovarian Cancers: Preliminary Results*

Fatime OUMAR DJIBRILLAH
*Department od Biomedical Engineering*
*Institute of Science and Engineering*
*Erciyes University*
*Kayseri, 38039, TURKEY.*
foumar855@gmail.com

Mehmet Emin YUKSEL
*Department of Biomedical Engineering*
*Faculty of Engineering*
*Erciyes University*
*Kayseri, 38039, TURKEY*
yuksel@erciyes.edu.tr

***Abstract*—Endometrial and ovarian cancers are the most common gynecological cancers. Early and accurate diagnosis of both diseases is essential for minimizing mortality rate as well as treatment costs. In this study, we propose a novel methodology for efficient detection of endometrial and ovarian cancers. The proposed approach is based on processing urine-based Fourier Transform infrared spectroscopy data by using machine learning powered by evolutionary feature selection. We propose a novel spectral feature selection method based on a binary genetic algorithm, which is combined with a number of well-known machine learning methods and applied on spectroscopy data obtained from urine samples. Our results show that the presented approach offers superior performance over other methods in the literature for the same purpose. It can therefore efficiently be used for accurate and automated detection of endometrial and ovarian cancers by utilizing a very small set of spectral features, which provides a significant benefit regarding physical implementation of the proposed pipeline as a biosensor kit.**

***Keywords—Endometrial cancer; Ovarian cancer; Spectral feature selection; Binary Genetic algorithm; Machine Learning; Classification.***

## I. INTRODUCTION

Cancer is one of the most common fatal diseases worldwide. According to the World Health Organization (WHO), cancer is the second leading cause of death after cardiovascular diseases and accounted for nearly 10 million deaths in 2020, roughly one in six deaths [1]. Endometrial cancer (EC) and ovarian cancer (OC) are the 6th and 8th most commonly diagnosed cancers in women [2], respectively. They are also the most common gynecological cancers.

Early diagnosis of both diseases is essential to reduce the mortality rate due to them. The standard diagnosis techniques for EC include ultrasound imaging, endometrial biopsy (which is considered as the Gold-standard diagnosis technique), hysteroscopy and sometimes Magnetic Resonance Imaging (MRI) [3]. Although these techniques generally offer good diagnostic accuracy, they also suffer from a number of drawbacks such as poor specificity, invasiveness, high economical cost, so on [3]. OC is conventionally diagnosed using serum cancer antigen 125 (CA 125), Human Epididymis Protein 4 (HE4) and transvaginal ultrasonography [4]. Some of these techniques are not sufficiently sensitive and specific for early diagnosis of OC while others cause considerable discomfort to the patient [4]. Therefore, accurate, cost-effective, and non-invasive diagnostic tools are highly desirable for the early diagnosis of both diseases and avoid the problems encountered with the aforementioned diagnosis methods.

Vibrational spectroscopy techniques, such as mid-Infrared (IR), Fourier Transform Infrared (FTIR) and Raman spectroscopy, have proven efficient over the past decade to provide discrimination between cancerous and healthy samples, demonstrating a promising role in both screening and diagnosis of cancers [3]. Previous research has shown that the diagnosis of several types of cancers including gynecological cancers [5,6,7] can efficiently be accomplished by analyzing biofluids such as blood, urine, etc., using different spectroscopic methods combined with appropriate machine learning (ML) methodologies.

Feature selection (FS) is a preprocessing technique commonly adopted in classification tasks to reduce the number of features employed in the classification process by selecting a subset of more relevant features from the original feature set. This process is sometimes used to improve the performance of the classifier and sometimes to allow the classifier to yield the same performance with a reduced number of features.

Motivated by these facts, we propose in this study a novel methodology for the accurate detection of EC and OC. The proposed approach is based on processing urine-based FTIR spectroscopy data by using ML powered by evolutionary feature selection. We propose a novel feature selection method based on a binary genetic algorithm, which is specifically tailored for spectral feature (wavenumber) selection in spectroscopic data.

978-1-6654-5432-2/22/$31.00 ©2022 IEEE

This new feature selection method is combined with a number of well-known ML methods and applied on urine-based spectroscopy data provided by [5]. Our results show that the presented approach offers superior performance over other methods in the literature. It can therefore efficiently be used for accurate and automated detection of EC and OC by utilizing a very small set of spectral features extracted from spectroscopic data coming from urine samples. This provides a significant benefit regarding physical implementation of the proposed pipeline as a biosensor kit.

## II. Related Works

A good number of researchers have proposed several methods for the automated cancer diagnosis using feature selection and ML approaches. For example, a comparison between the Attenuated Total Reflection Fourier Transform Infrared (ATR-FTIR) spectroscopy of biofluids and the currently used diagnostic modalities for OC detection was performed in [7]. Principal Component Analysis (PCA) was used as the feature selector and Discriminant Analysis (DA) was used as the classifier. The aim of this study was firstly to test the performance ATR-FTIR spectroscopy on biofluids for OC detection, and secondly to determine which biofluid was more accurate.

In a study performed in [8], feature selection based on GA to classify breast cancer was proposed. First, GA was used as a classifier to select the salient features from the dataset. Then several ML classifiers were used to classify whether the breast cancer was benign or malignant. Four different datasets were used and the results showed that the classifier trained with only 3, 5, 9 and 14 selected features offered good accuracy.

In [9], a new method for prostate cancer screening and classification based on Fourier Transform mid-infrared (FT-MIR) spectroscopy data, PCA, Successive Projections Algorithm (SPA), and GA. PCA, SPA, and GA were used for feature (wavenumber) reduction and selection while SVM was used as the classifier. Their results were compared with conventional SVM classifier (without variable selection) and showed that variable selection methods followed by SVM classifier increased the performance. GA-SVM exhibited a higher accuracy.

Mutual Informative GA (MI-GA) selection approach was applied to select informative gene in cancer data in [10]. MI was applied to select only the genes that have higher information related to cancer. Then the selected genes were given to GA in order to select the optimal set of genes required for accurate classification. SVM was used as the classifier in this study. The proposed method was applied to colon, lung, and ovarian cancer datasets and the results showed that the proposed gene selection was accurate.

## III. Material and Method

### A. Spectral Analysis (FTIR) and Dataset

FTIR spectroscopy is a technique that uses the beam of infrared radiations to identify the functional groups in a material [11]. In the FTIR technique, the interferograms of sample signal are obtained by using an interferometer and then they are transformed to IR spectra through Fourier Transform [12].

In this study, the ATR-FTIR spectroscopy data provided by [5] is used. They have been collected at the Royal Preston Hospital UK. The data set consists of FTIR spectroscopy samples of urine taken from patients with EC/OC as well as healthy individuals. There are in total 30 samples and 10 spectra were taken from each urine sample. In order to perform binary classification, the EC and OC are separated. The detail of the data set is summarized in Table 1.

Table 1. Details of the dataset.

| Dataset | Number of features | Number of instances |
|---|---|---|
| **EC** | 235 | 200 |
| **OC** | 235 | 200 |

### B. Feature Selection

FS is a process of reducing the number of features by selecting more relevant features to improve the performance of the classifier. The FS approach proposed in this study is performed by using an improved Binary Logic-based Genetic Algorithm (BL-GA) with a novel crossover operator based on logic OR operation. The details will be described in the next subsection.

### C. Genetic Algorithm

The Genetic Algorithm (GA) is a heuristic search method that is inspired by the theory of evolution which assumes the "survival of the fittest" [13]. GA consists of three main genetic operators including selection, crossover and mutation. In this study, we have aimed to choose salient spectral features and increase the classifier performance. In order to achieve this goal, a novel GA is proposed. The flow diagram of the proposed algorithm is shown in Figure 1.

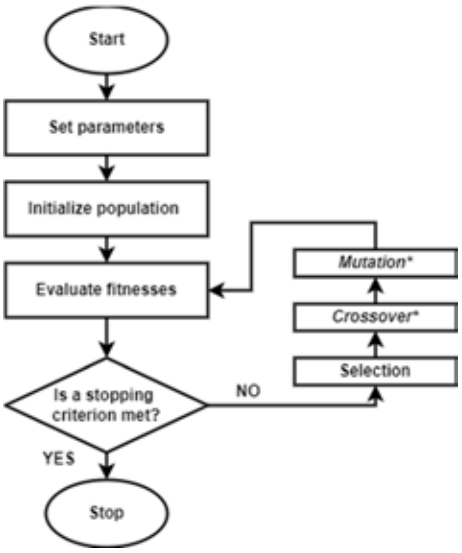


Fig. 1. Flow diagram of the algorithm.

Our proposed algorithm comprises the following steps:

1) The value of parameters that control the algorithm are set. These include parameters such as population size, maximum number of iterations, number of features to be selected, etc.
2) The initial population is generated. The population is made up of chromosomes and each chromosome is a vector of length equal to the total number of spectral features. Each element (gene) in each chromosome is either 1 or 0. A 0 in position k of a chromosome means that the kth feature is not selected while 1 means it is selected. Hence the total number of 1's in each chromosome represents the number of selected features.
3) After the chromosomes are generated, the fitness of each one is evaluated. The fitness evaluation function proposed in this study uses different ML classifiers including K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Naïve Bayes (NB). The classifiers are tested with 5-fold cross validation. Readers who want to get more information about these classifiers can refer to [14].
4) After the fitness is evaluated, the stopping criteria are checked. If a stopping condition has arisen the algorithm terminates, otherwise it continues its evolution.
5) The next generation is constructed. This is accomplished in three stages:
    a. Selection of 2 distinct chromosomes from the current population is performed based on Roulette selection.
    b. The selected parents are crossed over to generate an offspring. In this study we propose a novel crossover operator based on logic operation. The two parent chromosomes are logic-ORed with each other to generate the offspring. An example of this process is shown in Figure 2.

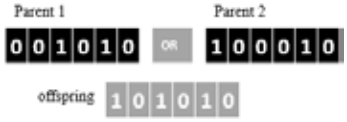


Fig. 2. Crossover operation.

    c. The offspring in general has more logic 1's than needed. Therefore, as a mutation operation, a desired number of logic 1's are randomly chosen and preserved while the others are cleared to zero.
6) After generating the new chromosomes, the fitness of each one is evaluated in step 3 and the process is repeated until one of the stopping criteria is met.

## IV. EXPERIMENTAL RESULTS AND DISCUSSION

In order to evaluate the performance of the proposed method, three main metrics are computed, which are accuracy, sensitivity and specificity. Accuracy is the percentage of correctly classified samples, whether positive or negative. Sensitivity is the percentage of correctly classified positive samples while specificity is the percentage of correctly classified negative samples.

These metrics are calculated using the following equations:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100 \quad (1)$$

$$Sensitivity = \frac{TP}{TP+FN} * 100 \quad (2)$$

$$Specificity = \frac{TN}{TN+FP} * 100 \quad (3)$$

where TP is true positive, TN is true negative, FP is false positive and FN is false negative.

Readers who are interested in getting more information about these metrics may visit [15].

The proposed method for spectral feature selection is applied using an improved GA based on logical crossover operation. From 235 features, only 8 features are identified by BL-GA to be more relevant. After getting the best subset, classification algorithms mentioned earlier in this section are applied to evaluate their performances.

Table 2 and Table 3 list the results for the comparative performances of the classifiers without and with the proposed FS method.

Table 2. The accuracy, sensitivity and specificity of classifiers without and with the proposed FS for EC detection.

| | EC detection without FS | | | EC detection with FS (Proposed method) | | |
|---|---|---|---|---|---|---|
| | ACC | SPEC | SENS | ACC | SPEC | SENS |
| KNN | 96.48 | 94.45 | 98.57 | **99.60** | **99.27** | **99.80** |
| SVM | 98.52 | 98.21 | 98.80 | **99.70** | **99.80** | **99.53** |
| NB | 80.96 | 84.54 | 78.16 | **92.70** | **95.20** | **90.04** |

Table 3. The accuracy, sensitivity and specificity of classifiers without and with the proposed FS for OC detection.

| | OC detection without FS | | | OC detection with FS (Proposed method) | | |
|---|---|---|---|---|---|---|
| | ACC | SPEC | SENS | ACC | SPEC | SENS |
| KNN | 96.23 | 97.53 | 94.88 | **99.99** | **100** | **99.81** |
| SVM | 99.32 | 99.45 | 99.19 | **99.35** | 98.86 | **99.76** |
| NB | 76.03 | 88.29 | 69.76 | **93.90** | **96.91** | **91.10** |

The performance metrics listed in both of these tables clearly demonstrate that the proposed FS method significantly improves the performances of all classifiers for the detection of both EC and OC.

In order to further reveal the strength of the proposed method, its diagnosis performance is compared with a number

of representative methods from the literature for EC and/or OC detection based on FTIR spectral analysis, results of which are listed in Table 4. These results provide an additional confirmation for the superiority of the diagnostic performance of the proposed method over other existing methods used for the same purpose.

## V. CONCLUSION

In this study, a novel feature selection method based on binary GA is proposed to choose salient FTIR spectral features for an accurate diagnosis of Endometrial and Ovarian cancers. The KNN, SVM and NB classifiers are used to evaluate the performance of the proposed method. The results show that the proposed method can efficiently identify and select relevant spectral features and enable classifiers to better detect EC and OC than the other methods in the literature. Based on the findings presented in this paper, it is concluded that the proposed method can efficiently be used for accurate and automated detection of EC and OC by utilizing a very small set of spectral features extracted from spectroscopic data coming from urine samples.

Table 4. The proposed method compared with previous works for the same purpose.

| Methods | Cancers | Biofluid | Accuracy | Specificity | Sensitivity |
|---|---|---|---|---|---|
| GA-LDA [5] | Endometrial | Urine | 90.0 | 100 | 70.0 |
| | Ovarian | | 98.3 | 97.5 | 100 |
| PLS-DA [7] | Ovarian | Urine | 76 | 87 | 29 |
| | | Plasma | 81 | 84 | 71 |
| | | Blood | 94 | 98 | 76 |
| PLS-DA [6] | Endometrial | blood | --- | 78 | 87 |
| **Proposed method** | Endometrial | Urine | **99.70** | **99.80** | **99.80** |
| | Ovarian | | **99.99** | **100** | **99.81** |

## REFERENCES

[1] World Health Organization. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer. (Accessed on 10 August 2022).

[2] World Cancer Research Fund International.(published on 23 March 2022). [Online]. Available: www.wcrf.org. (Accessed on 10 August 2022).

[3] R. Schiemer, D. Furniss, S. Phang, A. B. Seddon, W. Atiomo, K. B. Gajjar "Vibrational Biospectroscopy: An Alternative Approach to Endometrial Cancer Diagnosis and Screening". *International Journal of Molecular Sciences*. Vol. 23, April 2022.

[4] D. Žilovič, R. Čiurlienė, R. Sabaliauskaitė, S. Jarmalaitė. "Future Screening Prospects for Ovarian Cancer". *Cancers*, Vol. 13, July 2021.

[5] M. Paraskevaidi, C. L.M. Morais, K. M. G. Lima, K. M. Ashton, H. F. Stringfellow, P. L. Martin-Hirsch, F.L. Martin. "Potential of mid-infrared spectroscopy as a non-invasive diagnostic test in urine for endometrial or ovarian cancer". *Analyst*, Vol. 143, pp. 3156-3163, June 2018.

[6] M. Paraskevaidi, C. L.M. Morais K. M. Ashton, H. F. Stringfellow, R. J. McVey, N. A. J. Ryan, H. O'Flynn, V.N. Sivalingam, S. J. Kitson, M. L. MacKintosh, A. E. Derbyshire, C. Pow, O. Raglan, , K. M. G. Lima, M. Kyrgiou, P. L. Martin-Hirsch, F.L. Martin, E. J. Crosbie. "Detecting Endometrial Cancer by Blood Spectroscopy: A Diagnostic Cross-Sectional Study." *Cancers*, May 2020.

[7] P. Giamougiannis, C.L.M. Morais, B. Rodriguez *et al.* "Detection of ovarian cancer (± neo-adjuvant chemotherapy effects) via ATR-FTIR spectroscopy: comparative analysis of blood and urine biofluids in a large patient cohort". *Anal. Bioanal. Chem.,* Vol. 413, pp. 5095–5107, June 2021.

[8] S. Singla, P. Ghosh, U. Kumari, "Breast Cancer Detection using Genetic Algorithm with Correlation based Feature Selection: Experiment on Different Datasets," *International Journal of Computer Sciences and Engineering*, Vol.7, pp. 406-410, April 2019.

[9] L.F.S. Siqueira, C.L.M. Morais, R.F. Araújo Júnior, A.A. de Araújo, K.M.G. Lima. "SVM for FT-MIR prostate cancer classification: An alternative to the traditional methods". *Journal of Chemometrics*, August 2018

[10] M. Jansi Rani, D. Devaraj, "Two-Stage Hybrid Gene Selection Using Mutual Information and Genetic Algorithm for Cancer Data Classification". *J. Med. Syst.,* Vol. 43**,** June 2019.

[11] S.A. Khan, S.B. Khan, L.U. Khan, A. Farooq, K. Akhtar, A.M. Asiri, Fourier Transform Infrared Spectroscopy: Fundamentals and Application in Functional Groups and Nanomaterials Characterization. Sharma, S. (eds.) *Handbook of Materials Characterization*. Springer, Cham, 2018.

[12] A. Erkahveci and A. Karaali , "Fourier Transform Infrared (FTIR) Spektroskopinin Gıda Analizlerine Uygulanması (İngilizce)", *Gıda*, vol. 21, no. 5, Oct. 1996.

[13] D. E. Goldberg, J. H. Holland, "Genetic algorithms and machine learning. Machine Learning", *Machine Learning*, Vol.2, pp. 95–99 , 1988.

[14] S. B. Kotsiantis, I. Zaharakis, P. Pintelas, et al. "Supervised machine learning: A review of classification techniques". *Emerging Artificial Intelligence Applications in Computer Engineering*, Vol. 160, no 1, pp. 3-24, 2007.

[15] W. Zhu, N. Zeng, N. Wang, "Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations". *NESUG proceedings: Health Care and Life Sciences, Baltimore, Maryland*, *19*, vol. 67, 2010.