# A New Hybrid Global Optimization Approach for Selecting Clinical and Biological Features that are Relevant to the Effective Diagnosis of Ovarian Cancer

Abeer Alzubaidi
College of Science and Technology
Nottingham Trent University
Nottingham, UK
abeer.alzubaidi022014@my.ntu.ac.uk

Georgina Cosma
College of Science and Technology
Nottingham Trent University
Nottingham, UK
georgina.cosma@ntu.ac.uk

David Brown
College of Science and Technology
Nottingham Trent University
Nottingham, UK
david.brown@ntu.ac.uk

A. Graham Pockley
John van Geest Cancer Research Centre
College of Science and Technology
Nottingham Trent University
Nottingham, UK
graham.pockley@ntu.ac.uk

***Abstract*— Reducing the number of features whilst maintaining an acceptable classification accuracy is a fundamental step in the process of constructing cancer predictive models. In this work, we introduce a novel hybrid (MI-LDA) feature selection approach for the diagnosis of ovarian cancer. This hybrid approach is embedded within a global optimization framework and offers a promising improvement on feature selection and classification accuracy processes. Global Mutual Information (MI) based feature selection optimizes the search process of finding best feature subsets in order to select the highly correlated predictors for ovarian cancer diagnosis. The maximal discriminative cancer predictors are then passed to a Linear Discriminant Analysis (LDA) classifier, and a Genetic Algorithm (GA) is applied to optimise the search process with respect to the estimated error rate of the LDA classifier (MI-LDA). Experiments were performed using an ovarian cancer dataset obtained from the FDA-NCI Clinical Proteomics Program Databank. The performance of the hybrid feature selection approach was evaluated using the Support Vector Machine (SVM) classifier and the LDA classifier. A comparison of the results revealed that the proposed (MI-LDA)-LDA model outperformed the (MI-LDA)-SVM model on selecting the maximal discriminative feature subset and achieved the highest predictive accuracy. The proposed system can therefore be used as an efficient tool for finding predictors and patterns in serum (blood)-derived proteomic data for the detection of ovarian cancer.**

***Keywords - Genetic Algorithm; Hybrid Feature Selection; Cancer Diagnosis; Mutual Information; Predictive Modelling; Linear Discriminant Analysis; Support Vector Machine***

## I. INTRODUCTION

Ovarian cancer is the fifth most common cause of cancer-related death in women. The disease is essentially asymptomatic until late stages, at which point the 5-year relative survival rate is <44%. If detected and treated early in its progression, the 5-year survival rate is ~90%. [1]. For this reason, early diagnosis can significantly decrease the morbidity and mortality rate from ovarian cancer and help medical experts to make important patient management decisions. Currently, statistical methods are still used for cancer predictive modelling in clinical practice. However, it is a challenging task for traditional methods to analyse high dimensional data and handle the uncertainty and imprecision which is found in clinical data. Furthermore, cancer datasets contain a lot of irrelevant and redundant features which are considered as noise and could degrade the efficiency of cancer prediction model.

Several researchers have investigated the problem of automatic diagnosis of different types of cancer in the past. Polat and Gunes [2] proposed a lung cancer prediction model which takes the results of various medical tests carried out on a patient as input. Their system used Principal Component Analysis (PCA) to reduce the dimensionality of the feature space to four dimensions, with a Fuzzy Weighting scheme being used before the classification step. The data are then classified using an Artificial Immune Recognition System. Cosma et al. [3] proposed a neuro-fuzzy model for predicting the pathological stage in patients with prostate cancer. Their results revealed that the neuro-fuzzy system outperformed a statistical nomogram which is commonly adopted by clinicians to predict cancer stage prior to the pre-operative stage. A hybrid automatic system for cancer diagnosis based on Genetic Algorithm and Fuzzy Extreme Learning machines (ELM) was proposed by Daliri [4], in which the Genetic Algorithm was used to reduce dimensional of feature space and the results were fed to ELM for the classification. Wu et al. [5] proposed an Artificial Neural Network (ANN) to evaluate six tumour markers groups. Lu et al. [6] have presented a feature selection algorithm for lung cancer diagnosis using a Genetic Algorithm based on a separability criterion. Avci [7] proposed an expert system for cancer diagnosis which uses the General Discriminant Analysis (GDA) method to reduce the dimensionality of the feature space to eight dimensions, and then uses Least Square Support Vector Machine (LS-SVM) for the classification stage. Alzubaidi et al. [8] proposed a hybrid feature selection approach to breast cancer diagnosis which combines a Genetic Algorithm (GA) with Mutual Information (MI) for selecting the best combination of cancer predictors with maximal discriminative capability.

Herein, we propose a novel (MI-LDA) feature selection approach based global optimization framework for classifying ovarian and non-ovarian cancers. The proposed method involves hybridizing a Genetic Algorithm with Mutual-Information and Linear Discriminant Analysis (MI-LDA) approaches for selecting the maximal discriminative feature subset, in order to achieve a high level of predictive accuracy. Although Mutual Information (MI) is a popular and effective technique for feature selection problems, most of the MI-based feature selection algorithms adopt a local searching strategy approach which typically generates suboptimal solutions. A global search strategy for MI-based feature selection is presented in order to effectively select features and avoid being trapped in local optima. MI is utilized to guide the search process in a Genetic Algorithm (GA) in such a way that those insignificant candidate subsets of features can be discarded with respect to more correlated features to the consecutive generations. Then, the highly discriminative feature subsets are used to train the LDA classifier and test the classification error rate. A Genetic Algorithm is utilised to optimise the search process with respect to an error rate of the LDA classifier in order to select the best predictors among all candidate models. When the classification error rate is minimized, the output knowledge of the classifier must be maximized. The feature subset that produces the minimum classification error rate is selected as the optimal subset. It is important here to consider that the large amount of computational effort to train LDA on each of these irrelevant subsets of features is avoided. This results in a better performance and requires less time to process the results. The proposed approach is verified with the ovarian cancer data from the FDA-NCI Clinical Proteomics Program Databank Web site[1], and the solutions found are used to train the two classifiers: SVM and LDA. Finally the performance of the system is evaluated using various measures: the classification accuracy (CA), the Area Under the ROC Curve (AUC), True Positive Rate (TPR, Sensitivity) and False Positive Rate (FPR, measured as 1-Specificity).

## II. Related Work and Background

Limiting the number of features has become a fundamental step for constructing an intelligent diagnosis system whilst maintaining an acceptable level of classification accuracy. Two major techniques can perform dimensionality reduction [9]: feature extraction and feature selection. Feature extraction involves linear or nonlinear transformation from the original feature space to a new lower-dimensional one. Feature extraction approaches such as Principal Component Analysis, (PCA) and Singular Value Decomposition (SVD) can be applied to create a reduced representation of the original data which is then used to create prediction models. However, the challenging task is to determine the optimal number of features to retain - this is also known as the curse of dimensionality [10]. Feature selection approaches are different to feature extraction approaches. The aim of feature selection approaches is to select the most significant features (i.e. predictors) from the original feature space according to some criterion. This work relies on feature selection strategies for finding compact and more correlated features that are useful for building highly accurate risk prediction models.

The subsections that follow describe the evaluation criterion and search strategy of feature selection, and the Mutual Information approach which aims to measure the discriminating ability of a feature subset to distinguish different class labels.

### A. *Evaluation Criterion of Feature Selection*

The success of the feature selection process mainly depends on considering two aspects: defining an appropriate evaluation criterion and designing an effective search strategy [11]. Different evaluation criterions for feature selection have been proposed over the years [12]: filter, wrapper, and embedded. These criteria can be grouped into classifier-dependent approaches (wrapper and embedded methods), and classifier-independent approaches (filter methods) [13].

*Wrapper methods* rely on a predetermined supervised learning model. The learning model is retrained each time, a new subset is selected and then evaluated based on the empirical error of the learned model using robust validation strategies. In contrast, *filter models* separate the feature selection process from the classification process, and select feature subsets that are independent of any learning algorithm. The evaluation of filter methods is usually based on analysing the intrinsic properties of the data [14]. *Embedded methods* incorporate the feature selection process as part of the training process in order to reduce the computational time required for reclassifying different subsets which is undertaken in wrapper methods. It is important to consider the primary factors that distinguish feature selection approaches for defining an appropriate evaluation criterion. Computational speed and the chance of overfitting: in terms of speed, filters are faster than hybrid methods and these are, in turn, faster than wrappers. In terms of overfitting, wrappers have higher learning capacity so are more likely to overfit than hybrid approaches, which in turn are more likely to overfit than filter methods [13]. For high dimensional data, embedded methods will likely outperform filter methods in generalisation error, while wrappers become more computationally unfeasible as the number of features increases. Despite the lower demands of filters, a major disadvantage of this approach is that it does not interact with the learning algorithms, and this results in a lower performance than wrappers. However, wrapper models require high computational costs, particularly for high dimensional data.

Therefore, in this work we adopt an intermediate solution by using a hybrid approach which jointly considers the classifier design and the feature subset selection. Hence, both the high accuracy of wrappers and the efficiency and generality of filters are achieved to some extent. Guyon et al. [15] proposed an embedded approach for Support Vector Machines. Maldonado

[1] https://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp

et al. [16] proposed a Support Vector Machine-recursive feature elimination (SVM-RFE) approach to feature selection. A feature is considered to be useful on the basis of its weight which results from training SVMs with a set of feature. Embedded approaches for feature selection problems that use Linear Discriminant Analysis (LDA) techniques have been proposed in [17], [18].

### B. *Search Strategy of Feature Selection*

Different search strategies for feature selection have been used to generate best feature subsets and to progress the search processes. The optimal solution for large finite spaces is computationally complicated due to the resulting exponential search space [19], [20], [21], since most classical search algorithms adopt a local search strategy to find sub-optimal and near optimal solutions. On the other hand, finding sub-optimal solutions is quite difficult for those search algorithms due to the involvement of partial search in the solution space or computational complexity [22]. As a consequence, interest in feature selection approaches has recently shifted towards the global search algorithms (or, Metaheuristic). Metaheuristic search strategies have been used to find an optimal solution to a given problem by searching a full space rather than partial space. Furthermore, from a computational perspective, the feature selection problem is generally difficult to solve. It is inherently a combinatorial optimization problem. Metaheuristic approaches are particularly suitable for solving multi-objective combinatorial optimisation problems [23] for which the objective function for each combination of features in large multi-dimensional spaces must be computed. Since the Genetic Algorithm is one of the most widely used optimization methods, it has been naturally employed to solve feature selection problems that are focussed on finding solutions in complex and nonlinear search spaces.

A study by Siedlecki and Sklansky [24] exposed evidence that the Genetic Algorithm had a solid capability to reduce the time for finding near optimal features from large sets compared to other algorithms. Oh et al. [19] proposed a hybrid algorithm for finding the better solutions in the neighbourhood of each solution found by the Genetic Algorithm. A comparison of algorithms that select features for pattern recognition was conducted concluded that Genetic Algorithms are best suited for large-sized problems [25]. Subsequently, there have been many publications demonstrating the advantages of Genetic Algorithms for feature selection problems [26], [27], [28].

Genetic Algorithms (GAs) are the main paradigm of evolutionary computing, and are a rapidly growing area of Artificial Intelligence. GAs are adaptive heuristic search algorithms which were invented by Holland in the 1960s, and inspired from Darwin's theory of evolution "survival of the fittest" [29], [30]. GAs comprise the evolution process based optimization problems techniques and have been successfully applied to optimization problems, including pattern recognition and classification tasks. The Genetic Algorithm starts with a population of individuals which are called Chromosomes. These are possible candidates for an optimization problem. Each Chromosome is evaluated on the basis of its fitness quality in order to survive to the next generation. Crossover and mutation are then used to recombine the strongest Chromosomes in order to enable adaptation to the external environment. Usually, such candidate solutions are encoded in a binary string of 0 and 1. In the binary string, 0 indicates that the associated features has not been selected and removed from feature set, whereas 1 shows that its corresponding feature has been selected. Three major factors can change how the optimization scheme is performed by GAs to solve particular problem. An objective function must be specified in order to evaluate the quality of each candidate solution, a representation for candidate solutions, and genetic operators and stopping criteria must be specified.

### C. *Mutual Information*

Efficient feature selection must consider the importance of the evaluation criteria for measuring classification performance. The aim of the evaluation function is to measure the discriminating ability of a feature subset in order to distinguish different class labels. Currently, correlation analysis measures such as Granger causality analysis, Pearson correlation coefficient, canonical correlation analysis (CCA) and mutual information (MI) are the most commonly used methods [31]. MI has attracted the most attention and is considered to be a good indicator of the correlation between features and class labels, and it is not sensitive to noise or outlier data.

Entropy, divergence and mutual information are basic concepts defined within information theory. In its origin, information theory was used within the context of communication theory to find answers about data compression and transmission rate. Since then, information theory principles have been largely incorporated into machine learning. The data classification process is aimed at reducing the amount of uncertainty or gaining information about the target (classification) attribute. In Shannon's information theory, information is defined as that which removes or reduces uncertainty [32]. For a classification task, more information means higher accuracy of a classification model, since the predicted class of new instances is more likely to be identical to their actual class. A model that does not increase the amount of information is useless and its predictive accuracy is not expected to be better than a random guess.

Mutual Information measures the statistical dependency between random variables. The MI-based feature selection algorithm [33] is used to maximise the joint MI (Maximal Dependency, MD) between the input features and target output to select the more correlated subset from the original space. For two random variables *X* and *Y* with probability density function *p(x)* and *p(y)* respectively, the MI is defined as equation (1). Where *p(x, y)* is the joint probability density function of *X* and *Y*.

$$I(X;Y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy \qquad (1)$$

To consider the dependency between multiple random attributes $\{X_1, X_2, \dots X_n\}$ and $Y$, we can obtain the joint MI as equation (2). Where $p(x_1, x_2, \dots x_n)$ and $p(x_1, x_2, \dots x_n, y)$ are joint probability density functions.

$$I(X_1,X_2,\dots,X_n; Y) = \iint \dots \int p(x_1, x_2, \dots x_n, y)$$

$$log \frac{p(x_1, x_2 \dots x_n, y)}{p(x_1, x_2 \dots x_n)p(y)} dx_1dx_2 \dots dx_ndy \qquad (2)$$

The joint MI maximises the correlation between feature and the target class $Y$, and also involves the internal correlation between features. Therefore, the joint MI is highly appropriate for solving feature selection problems. However, when applying the MD criteria for high dimensional data, the accuracy of the MI estimator gradually decreases and the computational complexity rapidly increases, and this consequently limits the applications of this method [31]. Therefore, many mutual information-based feature selection algorithms have moved towards low-dimensional MI [34]. The simplest approach for MI-based features selection is the Maximal Relevance (MR) criterion because it is easy to implement and has high computational efficiency. The concept of the MR criterion is to maximise the pairwise MI between feature and target class in order to select the more correlated features:

$$MR = \{I(X, Y)\} \quad (3)$$

Although many MI based feature selection approaches have been proposed [35], [36], [37], [38], most of these algorithms attempted to maximise the relevancy in greedy way. In this paper, MI based feature selection is applied to global search algorithms for cancer diagnosis and prognosis purpose.

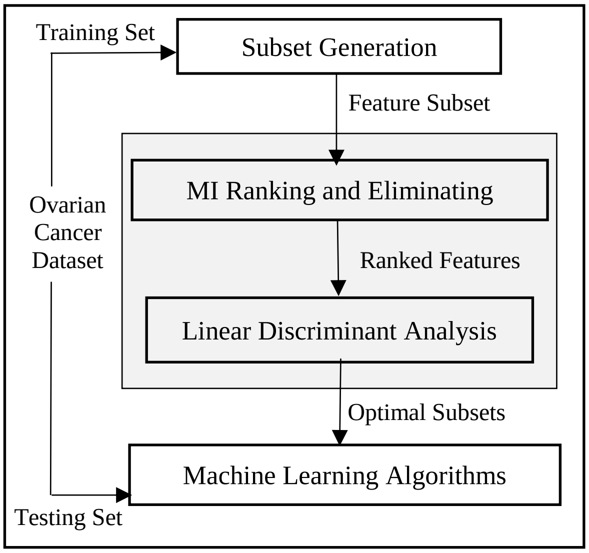

Fig. 1. The framework for the hybrid selection approach

## III. The proposed Method

A hybrid (MI-LDA) extraction approach to distinguish between patients with cancer and without cancer which is based on an evolutionary algorithm is proposed. The aim of the approach is multi-objective: 1) to reduce the number of features whilst keeping important correlations between features and the target class; and to 2) achieve a high predictive accuracy. The proposed model which combines the efficiency and accuracy of (MI-LDA) with the powerful global search ability of GA is shown in Fig. 1. The main steps of the GA with hybrid extraction applied to an ovarian cancer dataset is described in the following steps:

- **Step 1: Create initial population.** The algorithm begins by creating a random initial population. In this work, we set the initial population as an *m*-by-*n* matrix where the number of rows *m* represent the number of individuals (Chromosomes) in the population, which is equal to the value of population size in the population options. Population size determines the size of the population at each generation. The number of columns *n* is the number of genes (features) in each Chromosome (Genome) and that is exactly equal to the independent variables for the fitness functions.
- **Step 2: Objective Functions.** In order to converge to the optimal solution, two objective functions for the optimization algorithm are introduced:

1) **MI Ranking and Eliminating.** At the first objective function for the genetic algorithm, MI is used to maximize the inter-correlation between Chromosomes and the class attribute. MI is utilised to guide the search process in GA in order to select those individuals in the population based on their correlation with the target class. The best fit Chromosomes are then passed to the second optimisation function. GA and MI are successful in optimising the search process and reducing the feature space for the next optimization phase significantly.

2) **Linear Discriminant Analysis Model.** The second optimisation function adopts a Linear Discriminant Analysis (LDA) technique which is widely utilised in statistics, pattern recognition and machine learning to find a linear combination of features for classification, or for dimensionality reduction purposes. The candidate subsets of correlated features that are attributed to cancer diagnosis, which are extracted from the first optimisation function, are used to train the Linear Discriminant Analysis (LDA) classifier and test the classification error rates. The GA is performs the classifier dependent optimization in order to select the best predictors among all candidate models. The GA based LDA optimises the search process for the optimal subset of features quickly and precisely because the classifier has already been trained on highly correlated predictors. Furthermore, the irrelevant features have already been eliminated at the first optimisation stage, thereby leading to reduced computational effort and highly predictive results.

- **Step 3: Create new generation**. GAs use the individuals in the current generation to create the next population by choosing parents for the next generation on the basis of their fitness values from the objective functions. The probability for an individual to be selected is proportional to its minimum classification error rate value. Individuals with lower fitness have a better chance of surviving into the next generation. Some of the individuals in the current population that have best (lowest) fitness are chosen as elite. These elite individuals are passed to the next population. Individuals in the current generation are used to create the children (offspring) that form the next generation. Children are produced by making random changes to a single parent (mutation) and by combining the vector entries of a pair of parents (crossover). Both processes are essential to the genetic algorithm. Crossover enables the algorithm to extract the best genes (features) from different individuals and recombine them into potentially superior children. Mutation adds to the diversity of a population and thereby increases the likelihood that the algorithm will generate individuals with better fitness values.

- **Step 4:** Makes up the next generation by replacing the current population with the offspring.
- **Step 5:** Repeats step 2 to step 4 until the stopping criteria is met (maximum generation, mx, is reached). For the experiments, the value of mx was set to 50.
- **Step 6:** Selects the optimal subsets: The chromosome with the minimum classification error rate value in the final generation is selected as the optimal solution. After the evolution process, the selected subset of predictors is input into the LDA classifier or the SVM classifier for ovarian cancer diagnosis.

## IV. Experimental Methodology

This section describes the methodology and datasets that have been used for developing the strategies for diagnosing ovarian cancer. This study uses the Ovarian Cancer dataset, which is publicity available on the FDA-NCI Clinical Proteomics Program Databank website. This high-resolution ovarian cancer dataset was generated using the WCX2 protein array to identify serum (blood-derived) proteomic patterns that differentiate the serum of patients with ovarian cancer from that of women without ovarian cancer. It contains records collected from 216 samples with 4000 attributes. Each sample has one of two possible classes: Normal or Cancer. According to the class distribution, 121 (56%) instances were derived from patients with cancer and 95 (44%) instances were derived from women without cancer.

Several experiments were performed to demonstrate the ability of the hybrid feature selection approach to distinguish patients with ovarian cancer from those without. Classification models take as input a matrix A of size $m \times n$ where $m$ is the number of samples (patients) and $n$ is the number of attributes (i.e. features). Experiments were conducted using the state-of-art Leave-One-Out Cross Validation (LOOCV) approach. At the end of each experiment, the performance of the (MI-LDA)-LDA and (MI-LDA)-SVM models was compared to determine the least number of features which could be used to achieve highest predictive performance. The results of the comparison are shown in TABLE I.

To assess the performance of the approaches, we adopted various evaluation criterion. The results were evaluated using the Area Under the ROC Curve (AUC), classification accuracy (CA), True Positive Rate (TPR, Sensitivity) and False Positive Rate (FPR). Classification Accuracy (CA) refers to the percentage of correct classifications that are produced by a prediction model. The Receiver Operating Characteristic (ROC) can be used to establish a cut-off value for optimal performance of the system. AUC is used to differentiate between the data records in given classes (e.g. Cancer or Normal). The aim is to determine the cut-off point for which the classifier returns the high number of true positives and the low number of false positives.

True Positive Rate (i.e. Sensitivity) measures the proportion of actual positives which are correctly identified as such (e.g. the percentage of patients with cancer who are correctly identified as having cancer). True Negative Rate (i.e. Specificity) measures the proportion of negatives which are correctly identified as such (e.g. the percentage of patients without cancer who are correctly identified as such). A perfect system would return 100% Sensitivity (e.g., all patients with cancer are correctly classified) and 100% Specificity (e.g. all patients without cancer are correctly classified).

TABLE I. Performance of the proposed (MI-LDA)-LDA and (MI-LDA) -SVM models

| **Evaluation Measures** | **(MI-LDA)-LDA** | **(MI-LDA)-SVM** |
|---|---|---|
| **No. of Features** | 9 | 11 |
| **CA (%)** | 100 | 100 |
| **AUC** | 1.00 | 1.00 |
| **FPR** | 0.00 | 0.00 |
| **TPR** | 1.00 | 1.00 |

## V. The Results and Discussion

The results presented in this section determine the true ability of a hybrid extraction system to discriminate patients with ovarian cancer from those without according to the knowledge which has been acquired by the model during the learning process. To perform these evaluations, the actual outputs (i.e. predicted diagnosis) returned by each model during the validation stage were compared against the targets class (i.e. known diagnosis) of the ovarian cancer dataset. The best system would return the largest AUC, highest classification accuracy, highest Sensitivity, and highest Specificity.

The (MI-LDA)-LDA model was evaluated on the ovarian cancer dataset using the LOOCV approach. The performance accuracy of this model is shown in TABLE I. The results show that the (MI-LDA)-LDA model achieved perfect results (FPR=0, TPR=1, AUC = 1, CA= 100%) when using 9 out of 4000 original features. The ROC curve for the (MI-LDA)-LDA model when using 9 features is presented in Fig 2.

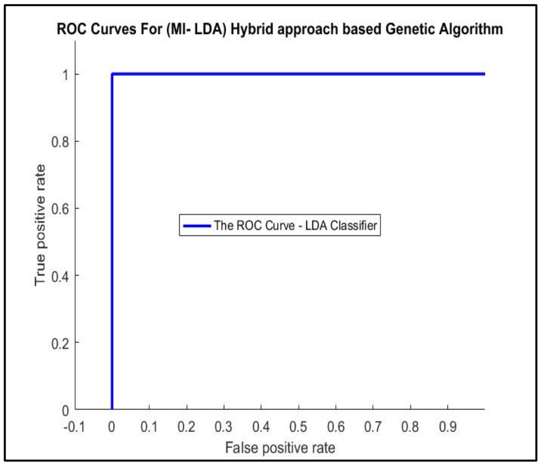

Fig. 2. ROC Curve for the (MI-LDA)-LDA model

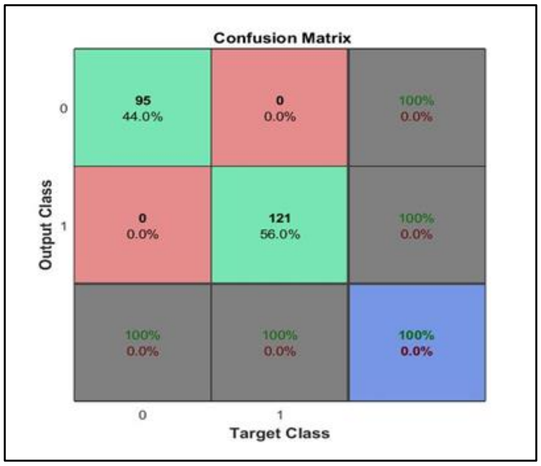

Fig. 3. Confusion matrix for the (MI-LDA)-LDA model

Fig. 3, shows the confusion matrix for the (MI-LDA)-LDA model. The first two diagonal cells show the number and percentage of correct classifications by the LDA. A total of 95 cases without ovarian cancer are correctly classified as such. This corresponds to 44.0% of all 216 examples. Similarly, 121 cases that were categorised as being individuals with ovarian cancer are correctly classified as such. This corresponds to 56.0% of all instances. Overall, 100% of the predictions are correct and 0.0% are incorrect classifications.

Fig. 4, shows the confusion matrix for the (MI-LDA)-SVM model. The Support Vector Machine classifier was trained using the linear kernel function. For the 95 cases without ovarian cancer, the predictions are 100% are correct and 0.0% are wrong. For the 121 cases that were categorised as being individuals with ovarian cancer are correctly classified as such. The first two diagonal cells show the number and percentage of correct classifications by the (MI-LDA)-SVM model. Overall, 100% of the predictions are correct and 0.0% are incorrect classifications.

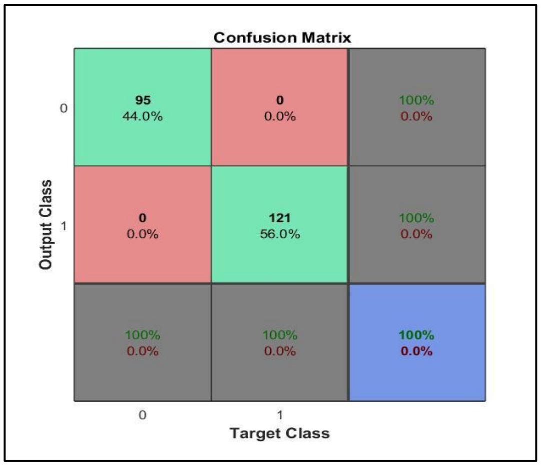

Fig. 4. Confusion Matrix for the (MI-LDA)-SVM model

The (MI-LDA)-SVM model was evaluated on the ovarian cancer dataset using the LOOCV approach. The performance accuracy of this model is shown in TABLE I. The ROC curve for (MI-LDA)-SVM is presented in Fig 5. The results show that (MI-LDA)-SVM model achieved prefect results when using 11 out of the 4000 features (FPR=0, TPR=1, AUC = 1, CA= 100%).

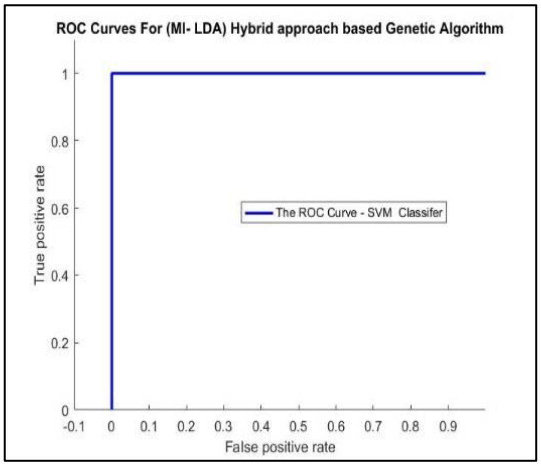

Fig. 5. ROC Curve (MI-LDA)-SVM model

From the above results, it can be concluded that the new (MI-LDA)-LDA hybrid approach which combines the global genetic searching and (MI-LDA) extraction approach is highly promising. The results show that the proposed model, (MI-LDA) feature selection approach with the LDA classifier (MI-LDA)-LDA required a fewer number of features (i.e. 9 features) to achieve 100% prediction accuracy, compared to the (MI-LDA) selection approach with linear SVM classifier (MI-LDA)-SVM, which required 11 features to achieve 100% prediction accuracy. The comparison results also showed considerable improvement in the computational time required to train the (MI-LDA)-LDA system as opposed to the (MI-LDA)-SVM model. These results suggest that the proposed (MI-LDA)-LDA model has great potential for finding predictors and patterns in serum proteomic data for the detection of ovarian cancer.

## VI. CONCLUSION

This paper explores the suitability of a hybrid framework which utilizes a Genetic Algorithm with a novel extraction (MI-LDA) approach to select the best set of features for identifying the presence of ovarian cancer using Machine Learning approaches. A Genetic algorithm is used to effectively select features and avoid being trapped in local optima where MI is used to maximize the MI between features and class labels. The highly discriminative features are passed to the Linear Discriminant Analysis (LDA) classifier to select the best features among all candidate models in which the Genetic Algorithm performs the classifier dependent optimization. The selected subset of features are then input into a classifier, such as Linear Discriminant Analysis (LDA) and Support vector machine (SVM). The results indicate that the proposed model, (MI-LDA)-LDA, which combines a hybrid extraction approach based global optimization algorithm for feature selection and then a Linear Discriminant Analysis for the risk prediction task (i.e. final classification) is a very promising approach for identifying patients with ovarian cancer. Future research includes experimenting with various high-dimensional datasets and more machine learning approaches, such as the Artificial Neural Network and the Naïve Bayes classifier.

## VI. REFERENCES

[1] D. A. Gaul, R. Mezencev, T. Q. Long, C. M. Jones, B. B. Benigno, A. Gray, F. M. Fernández, and J. F. Mcdonald, "Highly-accurate metabolomic detection of early-stage ovarian cancer," *Nat. Publ. Gr.*, vol. 30332, pp. 1–7, 2015.

[2] K. Polat and S. Güneş, "Principles component analysis, fuzzy weighting pre-processing and artificial immune recognition system based diagnostic system for diagnosis of lung cancer," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 214–221, 2008.

[3] G. Cosma, G. Acampora, D. Brown, R. C. Rees, M. Khan, and A. G. Pockley, "Prediction of Pathological Stage in Patients with Prostate Cancer: A Neuro-Fuzzy Model," *PLoS One*, vol. 11, no. 6, pp. 1–27, 2016.

[4] M. R. Daliri, "A Hybrid Automatic System for the Diagnosis of Lung Cancer Based on Genetic Algorithm and Fuzzy Extreme Learning Machines," *J. Med. Syst.*, vol. 36, no. 2, pp. 1001–1005, 2012.

[5] F. Feng, Y. Wu, Y. Wu, G. Nie, and R. Ni, "The Effect of Artificial Neural Network Model Combined with Six Tumor Markers in Auxiliary Diagnosis of Lung Cancer," *J. Med. Syst.*, vol. 36, no. 5, pp. 2973–2980, 2012.

[6] C. Lu, Z. Zhu, and X. Gu, "An Intelligent System for Lung Cancer Diagnosis Using a New Genetic Algorithm Based Feature Selection Method," *J. Med. Syst.*, vol. 38, no. 9, p. 97, 2014.

[7] E. Avci, "A new expert system for diagnosis of lung cancer: GDALS_SVM," *J. Med. Syst*, vol. 36, no. 3, pp. 2005–2009, 2011.

[8] A. Alzubaidi, G. Cosma, D. Brown, and A. G. Pockley, "Breast Cancer Diagnosis using a Hybrid Genetic Algorithm for Feature Selection based on Mutual Information," Interactive Technologies and Games (iTAG), 2016 International Conference on, Nottingham, 2016. To Appear.

[9] Z. M. Hira and D. F. Gillies, "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data," vol. 2015, no. 1, 2015.

[10] J. Paul, "Feature Selection from Heterogeneous Biomedical Data," Universit´e catholique de Louvain, 2015.

[11] J. Huang, Y. Cai, and X. Xu, "A hybrid genetic algorithm for feature selection wrapper based on mutual information," *Pattern Recognit. Lett.*, vol. 28, no. 13, pp. 1825–1844, 2007.

[12] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Nowé, "A survey on filter techniques for feature selection in gene expression microarray analysis.," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 9, no. 4, pp. 1106–19, 2012.

[13] G. Brown, A. Pocock, M.-J. Zhao, and M. Lujan, "Conditional Likelihood Maximisation: A Unifying Framework for Mutual Information Feature Selection," *J. Mach. Learn. Res.*, vol. 13, pp. 27–66, 2012.

[14] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, 2014.

[15] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[16] S. Maldonado, R. Weber, and J. Basak, "Simultaneous feature selection and classification using kernel-penalized support vector machines," vol. 181, pp. 115–128, 2011.

[17] R. Pique-Regi and A. Ortega, "Block Diagonal Linear Discriminant Analysis with Sequential Embedded Feature Selection," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, 2006, vol. 5, pp. V–V.

[18] L. Sheng, R. Pique-Regi, S. Asgharzadeh, and A. Ortega, "Microarray classification using block diagonal linear discriminant analysis with embedded feature selection," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 1757–1760.

[19] I.-S. Oh, J.-S. Lee, and B.-R. Moon, "Hybrid genetic algorithms for feature selection.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1424–1437, 2004.

[20] L. I. Kuncheva and L. C. Jain, "Nearest neighbor classifier : Simultaneous editing and feature selection," vol. 20, 1999.

[21] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," *Intell. Syst. their Appl. IEEE*, vol. 13, no. 2, pp. 44–49, 1998.

[22] M. L. Raymer, W. F. Punch, E. D. Goodman, L. a Kuhn, and K. Jain, "Dimensionality Reduction Using Genetic Algorithms," *IEEE Trans. Evol. Comput.*, vol. 4, no. 2, pp. 164–171, 2000.

[23] S. Kamyab and M. Eftekhari, "Neurocomputing Feature selection using multimodal optimization techniques," *Neurocomputing*, vol. 171, pp. 586–597, 2016.

[24] W. Siedlecki and J. Sklansky, "A note on genetic algorithms for large-scale feature selection," *Pattern Recognit. Lett.*, vol. 10, no. 5, pp. 335–347, Nov. 1989.

[25] M. Kudo and J. Sklansky, "Comparison of algorithms that select features for pattern classifiers," *Pattern Recognit.*, vol. 33, no. 1, pp. 25–41, 2000.

[26] J. K. F.J. Ferri, P. Pudil, M. Hatef, "Comparative Study of Techniques for Large-Scale Feature Selection," *Pattern Recognit. Pract. IV*, pp. 403–413, 1994.

[27] A. Jain and D. Zongker, "Feature Selection: Evaluation, Application, and Small Sample Performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 153–158.

[28] X. B. L. J.J. Liu, G. Cutler, W. Li, Z. Pan, S. Peng, T. Hoey, L. Chen, "Multiclass cancer classification and biomarker discovery using GA-based algorithms," *Bioinformatics*, vol. 21, no. 11, pp. 2691–2697, 2005.

[29] M. Melanie, "An introduction to genetic algorithms,"

*Cambridge, Massachusetts London, England*, p. 162, 1996.

[30] D. E. Goldberg and J. H. Holland, “Genetic Algorithms and Machine Learning,” *Mach. Learn.*, vol. 3, no. 2–3, pp. 95–99, Oct. 1988.

[31] M. Han and W. Ren, “Global mutual information-based feature selection approach using single-objective and multi-objective optimization,” *Neurocomputing*, vol. 168, pp. 47–54, 2015.

[32] J. R. Vergara and P. A. Este, “A review of feature selection methods based on mutual information,” pp. 175–186, 2014.

[33] P. A. Estvez, M. Tesmer, C. A. Perez, and J. M. Zurada, “Normalized mutual information feature selection,” *IEEE Trans. Neural Networks*, vol. 20, no. 2, pp. 189–201, 2009.

[34] Z. Wang, M. Li, and J. Li, “A multi-objective evolutionary algorithm for feature selection based on mutual information with a new redundancy measure,” *Inf. Sci. (Ny).*, vol. 307, pp. 73–88, 2015.

[35] H. H. Yang and J. Moody, “Data Visualization and Feature Selection: New Algorithms for Nongaussian Data,” *Adv. Neural Inf. Process. Syst.*, vol. 12, no. Mi, pp. 687–693, 1999.

[36] R. Battiti, “Using Mutual Information for Selecting Features in Supervised Neural Net Learning,” *IEEE Trans. Neural Networks*, vol. 5, no. 4, pp. 537–550, 1994.

[37] H. C. Peng, F. H. Long, and C. Ding, “Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, 2005.

[38] A. El Akadi, A. El Ouardighi, and D. Aboutajdine, “A Powerful Feature Selection approach based on Mutual Information,” *J. Comput. Sci.*, vol. 8, no. 4, pp. 116–121, 2008.