Breast Cancer Diagnosis using a Hybrid Genetic Algorithm for Feature Selection based on Mutual Information

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

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

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

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

***Abstract*— Feature Selection is the process of selecting a subset of relevant features (i.e. predictors) for use in the construction of predictive models. This paper proposes 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. The selected features are then input into a classifier to predict whether a patient has breast cancer. Using a publicly available breast cancer dataset, experiments were performed to evaluate the performance of the Genetic Algorithm based on the Mutual Information approach with two different machine learning classifiers, namely the k-Nearest Neighbor (K-NN), and Support vector machine (SVM), each tuned using different distance measures and kernel functions, respectively. The results revealed that the proposed hybrid approach is highly accurate for predicting breast cancer, and it is very promising for predicting other cancers using clinical data.**

Keywords— Genetic Algorithm; Feature Selection; Cancer Diagnosis; Mutual Information; Predictive Modelling

I. INTRODUCTION

Breast cancer is the most common cancer in women in both developed and developing countries where the number of breast cancer cases worldwide was estimated at 14.1 million new cases and 8.2 million deaths in 2012 [1]. Successful early detection can bring better treatments to patients, and can help medical experts make important decisions about patient healthcare. Statistical methods are the most popular approaches used in clinical practice for cancer diagnosis and prognosis. However, it is a challenging task for statistical methods alone to analyse high dimensional data, and to handle the uncertainly and imprecision which is typically apparent in clinical data.

Several methods are used to detect cancer in patients: such as blood tests, X-ray, CT scan, biopsy and patient examinations [2]. Data obtained from each test can hold important information which can be used by clinicians to better diagnose cancer and inform clinical decision-making. However, patient datasets contain a lot of irrelevant and redundant features, and the task is to select the features which are effective for a cancer prediction task. These features will be input into a cancer prediction model. Several researchers have investigated the problem of automatic diagnosis of different types of cancer in the past. Polat and Gunes [3] proposed an automatic diagnosis system to the task of diagnosing lung cancer. Their system uses Principal Component Analysis (PCA) to reduce the dimensionality of the feature space to four dimensions, and a Fuzzy Weighting scheme is used before the classification step. The data are then classified using an Artificial Immune Recognition System. A hybrid automatic system for cancer diagnosis based on Genetic Algorithm and Fuzzy Extreme Learning machines (ELM) was proposed by [2], where the Genetic Algorithm was developed to reduce the dimensionality of the feature space. The resulting features were input to ELM for performing the classification task. Wu et al. [4] proposed an Artificial Neural Network (ANN) to evaluate six tumor markers groups. Lu et al. [5] presented feature selection algorithm for detecting lung cancer using Genetic Algorithm based separability criterion. Avci [6] proposed an expert system for cancer diagnosis. Firstly, the General Discriminant Analysis (GDA) method was used to reduce the dimensionality of the feature space to eight dimensions, and then least square support vector machine (LS-SVM) was used in classification stage. Cosma et al. [7] 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 commonly adopted by clinicians to predict cancer stage prior at the pre-operative stage.

Our paper proposes a hybrid approach for identifying malignant from benign tumors. The proposed method is the combination of a Genetic Algorithm (GA) based on Mutual Information (MI) for solving feature selection problems. Most MI based techniques are greedy approaches, which usually generate suboptimal solutions. In this paper, the MI based feature selection approach is transformed into global optimization, where genetic searching is used to: 1) effectively select features and to avoid being trapped in local optima; and to 2) maximize the MI between features and class labels. The selected subset of features are then input in two classifiers: the Support Vector Machine (SVM), and the k-Nearest Neighbor (k-NN) classifier for performing the prediction task. The experiments were performed on the Wisconsin Diagnostic and

Prognostic Breast Cancer dataset[1]. The results show that only a subset of the features were required for reaching highest classification accuracy.

II. The Proposed method

Here we present an overview of the proposed hybrid selection approach for classifying between two types of tumors for breast cancer diagnosis, malignant and benign. This approach is illustrated in Fig. 1. Mutual Information (MI) is used to quantify the correlation between features and the target class and guide the Genetic Algorithm to selects those features that are more relevant for the diagnosis. The presented approach is evaluated using a publicly available Wisconsin breast cancer dataset, and the solutions found are used to train the two machine learning classifiers: Support Vector Machine (SVM) and k-Nearest Neighbour (k-NN). Finally the performance of the proposed approach is quantified using a number of evaluation measures: classification accuracy, Area Under the Curve (AUC), sensitivity and specificity.

A. Genetic Algorithms for Feature Selection

The main purpose of feature selection is to reduce the number of features used in classification while maintaining acceptable classification accuracy. Feature selection reduces the size of the data input into the prediction model; provides an understanding as to which features are required to allow for accurate differentiation between benign and malignant tumors; and can improve the performance of the prediction model [8]. Feature selection methods have therefore become an important step in clinical diagnosis systems.

The success of the feature selection process mainly depends on considering two aspects: search strategy and criteria [9]. Different feature selection approaches use different methods to generate subsets and progress the search processes. A number of comparative studies [10], [11], [12] have demonstrated that, for large finite spaces, finding the optimal solution is computationally expensive due to the resulting exponential search space. For this reason, most search strategies attempt to find sub-optimal and near optimal solutions [13]. Recent research interest has shifted toward 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 feature spaces. Metaheuristic algorithms are especially effective when the information is uncertain and dynamic. Therefore, these advanced approaches are efficient in dealing with biological predictive systems which require handling the uncertainly and imprecision which is apparent in clinical data and medical images. Since the Genetic Algorithm is one of the most widely used optimization methods for finding solutions in complex and nonlinear search spaces, it has been naturally employed to solve feature selection problems.

Genetic Algorithms are the main paradigm of evolutionary computing, and a rapidly growing area of Artificial Intelligence (AI). It turns out that there is no accurate definition of "Genetic Algorithms" formally accepted by the evolutionary computation community[14]. However, it can be said that Genetic Algorithms are adaptive heuristic search algorithm which are invented by Holland in the 1960s and inspired from Darwin's theory of evolution "survival of the fittest"[15]. Five important factors can change how the optimization scheme is performed by Genetic Algorithms (GA): population creation, fitness function, selection schema, genetic operators and stopping criteria. The algorithm starts with a population of binary strings which are called chromosomes. These are possible candidates for an optimization problem. During each iteration the populations are evaluated based on their fitness quality, and then the crossover and mutation genetic operators are utilised to select fitter solutions. Most commonly, the candidate solutions are encoded in a binary string of 0 and 1. In the binary string, 0 indicates that an associated feature has not been selected, whereas 1 shows that its corresponding feature has been selected.

A study by Siedlecki and Sklansky [16] revealed evidence that the Genetic Algorithm was faster in finding near optimal features from large datasets compared to other algorithms. Oh et al. [10] 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 in [17]. They concluded that Genetic Algorithms are best suited for large-sized problems. Subsequently, a lot of literature has been published which demonstrates the advantages of Genetic Algorithms for feature selection tasks [18],[19] [20].

B. Mutual Information (MI)

For efficient feature selection we must consider the importance of the evaluation criteria for measuring classification performance. All feature selection methods need to use an evaluation criteria together with a search strategy to obtain the optimal feature set. Among all of the evaluation criteria, mutual information has attracted the most attention because it is a good indicator of the correlation between features and class labels, and it is considered to be least sensitive to noise or outlier data than other approaches. The basic idea of the MI-based feature selection algorithm [21] is to select one optimal subset from the original dataset by maximizing the joint MI between the input features and target output. Estimation of high-dimensional MI is very difficult and has high computational complexity, and this consequently limits the applications of this method. Most of the existing methods adopt low-dimensional MI [22]. Many MI-based feature selection algorithms have been proposed [23],[24],[25],[26]. Most of these algorithms adopt suboptimal searching methods. In this paper MI based feature selection is shifted to global search algorithms for breast cancer diagnosis.

The main steps of the Genetic Algorithm with MI applied to the breast cancer dataset is described in the following steps:

- Step 1: Create initial population. The generation of the initial population is straightforward and created by setting each bit of the Chromosome to 0 or 1 randomly. The number of strings in the Chromosome equals to *n*, where n is the total number of features of the breast cancer dataset.
- Step 2: Compute the MI between features and target classes. In the main loop, each Chromosome in the population is evaluated based on the correlation with the

[1] Available from: https://archive.ics.uci.edu/ml/datasets.html

class. The chromosomes with the highest MI values, i.e. correlations with the target class are used in step3.

- Step 3: Fitness proportional selection is used for reproduction. The probability for a Chromosome to be selected is proportional to its MI value.
- Step 4: Crossover and mutation operators reproduce some fitter chromosomes and generate a new chromosome (offspring) to be used in a next generation.
- Step 5: Steps 2 to 4 are repeated until the maximum generation (this has been experimentally set to 80) is reached.
- Step 6: The individual in the final generation with the best fitness value is selected as the optimal solution. Once the evolution process is complete, the selected features are input into the SVM and k-NN classifiers for the breast cancer diagnosis.

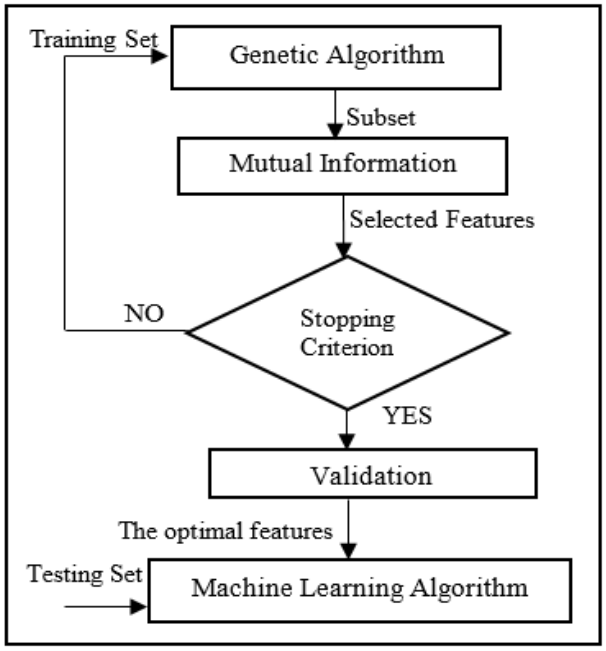

Figure 1. The overall schema of the GA-MI approach with machine learning classification

III. CLASSIFICATION

Once the feature selection process is complete, a subset of features returned is input into the classifier. The SVM (with different kernel functions) and K-NN classifiers (with different distance measures) were adopted.

A. Support Vector Machine (SVM) Classifier

Support Vector Machine (SVM) is a widely employed learning algorithm due to its superior data classification capability [5]. It can handle nonlinear classification problems by mapping the original training data to a high-dimensional feature space with a kernel function. It then determines the best separating hyperplane, which serves as a boundary separating the data from two classes. A binary classifier was trained to predict between two diagnosis classes: Malignant and Benign. This hyperplane should maximize the margins known as the distances between the nearest training points. In this study four kernel functions have been used. These kernel functions were the Radial Basis Function (RBF), Linear, Quadratic and Multi-Layer Perceptron (MLP).

B. k-Nearest Neighbour (k-NN) Classifier

The k-NN model adopts the nearest distance in finding the class label of new data in the training set. The feature space is separated to several regions during the training stage. Then these training samples are mapped into regions according to the similarity between the samples. Similarity between samples is computed using a distance measure, and this measure is usually selected experimentally. The outcome (i.e. predicted diagnosis) of a new patient case is predicted by taking into consideration the diagnosis outcomes of its k closest neighbours. The k closest cases are calculated using a distance measure. For the experiments described in this paper the value of *k* was set to 2 nearest neighbours, and this value was selected experimentally. Experiments were also conducted with four different types of distance measures. These are the Euclidean, Minkowski, Seuclidean and Correlation distance measures.

IV. EXPERIMENTAL METHODOLOGY

This section describes the methodology and dataset that have been used for breast cancer diagnosis experiments. In this paper, a breast cancer dataset was utilised to classify two types of tumors: benign and malignant.

A. Dataset

This study used the Wisconsin Breast Cancer dataset, which is publicity available on the UCI Machine Learning Repository website. The dataset is provided by university of Wisconsin hospital, Madison from Dr. William H. Wolberg [27]. It contains records collected from 699 patients. Each record contains 10 features including the diagnosis feature (i.e. known labels: malignant or benign), as shown in Table I. According to the class distribution 458 (65.5%) cases were derived from patients with a benign tumor and 241 (34.5%) cases were derived from patients with a malignant tumor. The patient samples consist of visually assessed nuclear features of fine needle aspirates (FNAs) taken from patients' breasts. Each feature (except the diagnosis feature) is in the interval 1 to 10, with value 1 corresponding to a normal state and 10 to a most abnormal state. The Diagnosis feature holds values 0 and 1, where 0 denotes a benign tumor diagnosis, and 1 denotes a malignant tumor diagnosis. Malignant tumor diagnosis is determined by taking a sample tissue from the patient's breast and performing a biopsy on it. A benign diagnosis is confirmed either by biopsy or by periodic examination, depending on the patient's choice.

B. Evaluation Measures

Several experiments have been performed to evaluate the ability of the proposed method for classifying between two types of tumors: benign and malignant. Classification models take as input a matrix A of size m×n where m is the total number of patient records and n is the total number of clinical features. Experiments were conducted using the state-of-art Leave-One-Out Cross Validation (LOOCV) approach and evaluation measures widely used in machine learning experiments on clinical data. The experiments were run using different number

of features, *1,…n*. At the end of each experiment, the performance of the SVM and k-NN classifiers was compared to determine the least number of features which could be used to achieve best system performance.

TABLE I. BREAST CANCER DATASET FEATURES

	Feature name	**Range**
1	Clump thickness	1-10
2	Uniformity of cell size	1-10
3	Uniformity of cell shape	1-10
4	Marginal adhesion	1-10
5	Single epithelial cell size	1-10
6	Bare nuclei	1-10
7	Bland chromatin	1-10
8	Normal nucleoli	1-10
9	Mitoses	1-10
10	Diagnosis	0 for benign, 1 for malignant.

To assess the performance of the proposed approach, we adopted various evaluation measures: classification accuracy (CA), the area under the ROC curve (AUC), and the Optimal ROC points (OPR): True Positive Rate (TPR, Sensitivity) and False Positive Rate (FPR, measured as 1-Specificity). Classification accuracy (CA) refers to the percentage of correct classifications produced by the trained k-NN and SVM classifiers on the testing set.

The Receiver Operating Characteristic (ROC) can be used to establish a cut-off value for optimal performance of the system (i.e. Optimal ROC points (ORP)). The area under the ROC curve (AUC) is used to differentiate between the data records in given classes (e.g. malignant or benign). The aim is to determine the cutoff point for which the classifier returns the highest number of true positives and the low number of false positives. Sensitivity (i.e. True Positive Rate) measures the proportion of actual positives which are correctly identified as such (e.g. the percentage of malignant tumors which are correctly identified malignant). Specificity (i.e. True Negative Rate) measures the proportion of negatives which are correctly identified as such (e.g. the percentage of benign tumors which are correctly identified as benign). A perfect system would return 100% sensitivity (e.g., all patients with malignant tumor are correctly classified) and 100% specificity (e.g. all patients with benign tumors are correctly classified).

C. Results and Discussion

1) **Experimental results using the GA-MI feature selection approach with the SVM classifier:** The results presented in this section determine the true ability of a system to discriminate malignant from benign tumors according to the knowledge which has been acquired by the system during the learning process. To perform these evaluations, the actual output (i.e. predicted diagnosis) returned by a model during the validation stage was compared against the target value (i.e. known diagnosis). The best system would return the largest AUC, a high Sensitivity (i.e. True Positive Rate), and a high Specificity (i.e. True Negative Rate) at ORPs. The SVM was trained using the Radial Basis Function (RBF), Linear, Quadratic, and the Multilayer Perceptron (MLP) kernel functions. The results of testing the performance of the SVM using the various kernel functions are shown in Table II. The comparison of the classification accuracy of SVM when using different kernel functions are illustrated in Fig. 2. The ROC curves for the four kernel functions are presented in Fig. 3. The results show that the Linear-SVM achieved the largest AUC (AUC = 0.9702, Correct Rate=0.9845) when using 7 features. However, the AUC was also very high, (AUC=0.9669, Correct Rate=0.9844) when using only 5 features and the Quadratic-Linear kernel function. The Quadratic-Linear kernel function is considered a more reliable kernel function to use, as more than often non-linear separation is needed.

TABLE II. SVM CLASSIFIEER RESULTS WITH VARIOUS KERNEL FUNCTIONS

Eval. Measures	**SVM - Kernel Functions**			
	RBF	***Linear***	***Quadratic***	***MLP***
2 Features				
Correct Rate	0.9575	0.9796	0.9853	0.9578
AUC	0.9278	0.9529	0.9264	0.9311
ORP FPR	0.0788	0.0373	0.0249	0.0788
ORP TPR	0.9345	0.9432	0.8777	0.9410
3 Features				
Correct Rate	0.9602	0.9669	0.9627	0.9559
AUC	0.9365	0.9470	0.9440	0.9323
ORP FPR	0.0747	0.0622	0.0705	0.0830
ORP TPR	0.9476	0.9563	0.9585	0.9476
4 Features				
Correct Rate	0.9753	0.9800	0.9843	0.9907
AUC	0.9510	0.9639	0.9636	0.9590
ORP FPR	0.0456	0.0373	0.0290	0.0166
ORP TPR	0.9476	0.9651	0.9563	0.9345
5 Features				
Correct Rate	0.9820	0.9822	0.9844	0.9795
AUC	0.9605	0.9659	0.9669	0.9508
ORP FPR	0.0332	0.0332	0.0290	0.0373
ORP TPR	0.9541	0.9651	0.9629	0.9389
6 Features				
Correct Rate	0.9778	0.9823	0.9844	0.9683
AUC	0.9607	0.9681	0.9669	0.9382
ORP FPR	0.0415	0.0332	0.0290	0.0581
ORP TPR	0.9629	0.9694	0.9629	0.9345
7 Features				
Correct Rate	0.9822	0.9845	0.9800	0.9909
AUC	0.9648	0.9702	0.9617	0.9688
ORP FPR	0.0332	0.0290	0.0373	0.0166
ORP TPR	0.9629	0.9694	0.9607	0.9541
8 Features				
Correct Rate	0.9887	0.9844	0.9584	0.9885
AUC	0.9678	0.9691	0.9387	0.9591
ORP FPR	0.0207	0.0290	0.0788	0.0207
ORP TPR	0.9563	0.9672	0.9563	0.9389
9 Features				
Correct Rate	0.9865	0.9823	0.9606	0.9909
AUC	0.9646	0.9670	0.9419	0.9688
ORP FPR	0.0249	0.0332	0.0747	0.0166
ORP TPR	0.9541	0.9672	0.9585	0.9541

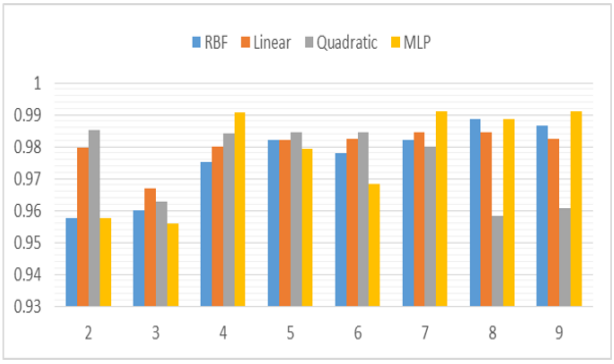

Figure 2. SVM Performance comparison using various kernel functions

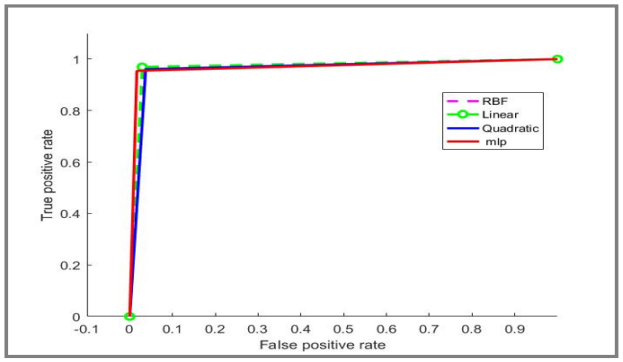

Figure 3. ROC Curves from SVM performance using various kernel functions

2) Experimental results using the GA-MI feature selection approach with the k-NN classifier: Several k-Nearest Neighbour (k-NN) models were created using various distance measures: Correlation, Minkowski, Euclidean and Seuclidean. The evaluation results of testing the performance of the k-NN using the various distance measures are presented in Table III. The comparison of classification accuracy of k-NN with various distance measures are illustrated in Fig. 4. The ROC curves when using various distance measures of k-NN are presented in Fig. 5. The results revealed that a high AUC is achieved (AUC=0.9678, Correct Rate=0.9887) when using 7 features and setting the k-NN with the Minkowsky and Euclidian distance measures. Although, at 7 features, k-NN achieved a slightly higher AUC (AUC=0.9679) when using the Seuclidean distance measure than when using the Minkowsky (AUC=0.9678) and Euclidean (AUC=0.9678) distance measures, the Correct Rate when using Seuclidean was lower (Correct Rate=0.9865) and for this reason the Minkowsky and Euclidian distance measures are better choices. Furthermore, observing the performance of the k-NN using the various distance measures, it appears that Minkowsky and Euclidian distance measures performed consistently better than the Correlation and Seuclidean distance measures. In addition, when using 8 features performance of k-NN decreases before it slightly increases again when using 9 features. Therefore, reducing the number of features to 7 would give highly accurate results.

In overall, the results suggest that predictive modelling accuracy was better when using the k-NN classifier and the proposed GA-MI feature selection approach.

TABLE III. K-NN CLASSIFIEER RESULTS WITH VARIOUS DISTANCE MEASURES

Eval. Measures	**k-NN Distance Measures**			
	Correlation	***Minkowski***	***Euclidean***	***Seuclidean***
2 Features				
Correct Rate	0.9380	0.9619	0.9619	0.9618
AUC	0.8322	0.9331	0.9331	0.9320
ORP FPR	0.7598	0.0705	0.0705	0.0705
ORP TPR	0.7598	0.9367	0.9367	0.9345
3 Features				
Correct Rate	0.8265	0.9759	0.9759	0.9784
AUC	0.7126	0.9214	0.9214	0.9257
ORP FPR	0.7074	0.0415	0.0415	0.0373
ORP TPR	0.7074	0.8843	0.8843	0.8886
4 Features				
Correct Rate	0.9276	0.9733	0.9733	0.9712
AUC	0.8691	0.9533	0.9533	0.9523
ORP FPR	0.8668	0.0498	0.0498	0.0539
ORP TPR	0.8668	0.9563	0.9563	0.9585
5 Features				
Correct Rate	0.9463	0.9843	0.9843	0.9798
AUC	0.8944	0.9636	0.9636	0.9584
ORP FPR	0.8843	0.0290	0.0290	0.0373
ORP TPR	0.8843	0.9563	0.9563	0.9541
6 Features				
Correct Rate	0.9416	0.9865	0.9865	0.9821
AUC	0.8881	0.9646	0.9646	0.9627
ORP FPR	0.8799	0.0249	0.0249	0.0332
ORP TPR	0.8799	0.9541	0.9541	0.9585
7 Features				
Correct Rate	0.9605	0.9887	0.9887	0.9865
AUC	0.9156	0.9678	0.9678	0.9679
ORP FPR	0.9017	0.0207	0.0207	0.0249
ORP TPR	0.9017	0.9563	0.9563	0.9607
8 Features				
Correct Rate	0.9624	0.9843	0.9843	0.9844
AUC	0.9133	0.9658	0.9658	0.9680
ORP FPR	0.8930	0.0290	0.0290	0.0290
ORP TPR	0.8930	0.9607	0.9607	0.9651
9 Features				
Correct Rate	0.9559	0.9910	0.9910	0.9888
AUC	0.9104	0.9731	0.9731	0.9733
ORP FPR	0.8996	0.0166	0.0166	0.0207
ORP TPR	0.8996	0.9629	0.9629	0.9672

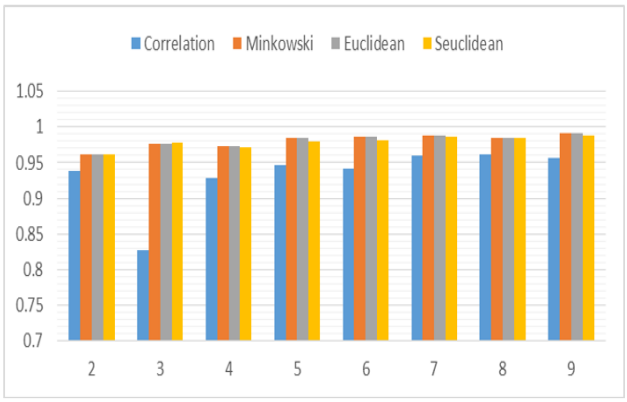

Figure 4. k-NN Performance comparison using various distance measures

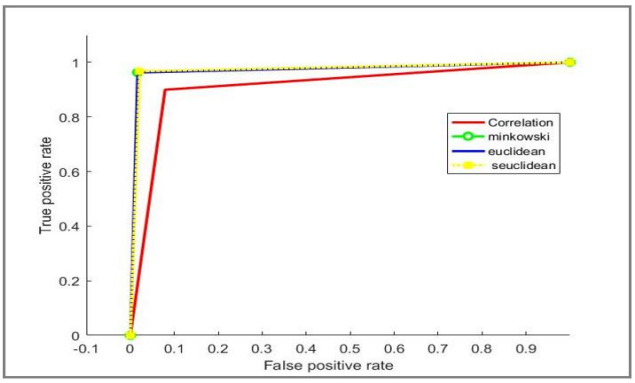

Figure 5. ROC Curves: k-NN performance using various distance measures

V. CONCLUSION

This paper explores the suitability of a hybrid framework which utilises a Genetic Algorithm with Mutual Information to select the optimal set of features for performing breast cancer prediction 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 selected subset of features was used as input for two classifiers, including k-Nearest Neighbour (k-NN), and Support vector machine (SVM). The results indicate that the proposed system can be very effective in cancer diagnosis and therefore it can be helpful for the physicians in this application.

REFERENCES

[1] S. Boughorbel, R. Al-Ali, and N. Elkum, "Model Comparison for Breast Cancer Prognosis Based on Clinical Data," *PLoS One*, vol. 11, no. 1, pp. 1–15, 2016.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

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

[9] M. M. Kabir, M. Shahjahan, and K. Murase, "A new local search based hybrid genetic algorithm for feature selection," *Neurocomputing*, vol. 74, no. 17, pp. 2914–2928, 2011.

[10] 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.

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

[12] 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.

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

[14] M. Melanie, "An introduction to genetic algorithms," *Cambridge, Massachusetts London, England, …*, p. 162, 1996.

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

[16] 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.

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

[18] and 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.

[19] 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.

[20] 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.

[21] P. A. Estévez, 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology.," *Proc. Natl. Acad. Sci.*, vol. 87, no. 23, pp. 9193–9196, 1990.