Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/bbe

Original Research Article

Determining relevant biomarkers for prediction of breast cancer using anthropometric and clinical features: A comparative investigation in machine learning paradigm

Bikesh Kumar Singh *

Department of Biomedical Engineering, National Institute of Technology, Raipur, Raipur, Chhattisgarh, India

ARTICLE INFO

Article history:
Received 4 September 2018
Received in revised form
9 March 2019
Accepted 13 March 2019
Available online 25 March 2019

Keywords:
Breast cancer biomarkers
Machine learning
Expert systems
Clinical features
Feature selection

ABSTRACT

Early detection of breast cancer plays crucial role in planning and result of associated treatment. The purpose of this article is threefold: (i) to investigate whether or not clinical features obtained using routine blood analysis combined with anthropometric measurements can be utilized for envisaging breast cancer using predictive machine learning techniques; (ii) to explore the role of various machine learning components such as feature selection, data division protocols and classification to determine suitable biomarkers for breast cancer prediction; and (iii) to evaluate a recent database of clinical and anthropometric measurements acquired from normal individuals and individuals suffering from breast cancer. A database consisting of anthropometric and clinical attributes is used in the experiments. Various feature selection and statistical significance analysis methods are used to determine the relevance of various features. Furthermore, popular classifiers such as kernel based support vector machine (SVM), Naïve Bayesian, linear discriminant, quadratic discriminant, logistic regression, *K*-nearest neighbor (*K*-NN) and random forest were implemented and evaluated for breast cancer risk prediction using these features. Results of feature selection techniques indicate that among the nine features considered in this study, glucose, age and resistin are found to be most relevant and effective biomarkers for breast cancer prediction. Further, when these three features are used for classification, the medium *K*-NN classifier achieves the highest classification accuracy of 92.105% followed by medium Gaussian SVM which achieves classification accuracy of 83.684% under hold out data division protocol.

* *Corresponding author at*: Department of Biomedical Engineering, National Institute of Technology, Raipur, G.E. Road, Raipur, Chhattisgarh 492010, India.

E-mail address: bsingh.bme@nitrr.ac.in

https://doi.org/10.1016/j.bbe.2019.03.001

1. Introduction

Breast cancer is a serious health issue as it accounts for more than 1.6% of the total deaths in females worldwide. As per the recent reports of American Cancer Society, in 2016, around 246,660 fresh cases of invasive breast cancer will be detected in women [1]. Screening techniques for breast cancer have demonstrated to be an imperative approach for declined death rates due to breast cancer in United States [2]. At present, mammography and ultrasound are commonly used breast cancer screening techniques. However, both these techniques have certain limitations which pose restrictions on performance of computer aided diagnosis (CAD) systems based on breast mammogram and ultrasound images. The general procedure of image based CAD systems include image acquisition along with patient demographic and clinical attributes, image normalization, image preprocessing for noise removal, lesion detection and segmentation, feature extraction, feature selection and classification [3–6]. However, such a procedure can be computationally expensive due to large number of steps involved. Further, inaccurate segmentation and irrelevant features can result in imprecise prediction models [7]. The alternative to this approach as reported by several researchers is to determine suitable biomarkers based on routine blood analysis and anthropometric attributes for breast cancer prediction. This article proposes a decision support system to assist clinicians in predicting the breast cancer based on clinical and anthropometric attributes of patients. Exhaustive experiments are conducted to evaluate and validate the proposed prototype.

1.1. *Related work*

This section presents some of the recent studies reported in the proposed research area. In Opstal-van Winden et al. [8], principal component analysis (PCA) and random forest (RF) analysis is employed for predicting breast cancer in early stages using a set of ten prospective serum biomarkers and cancer antigens collected prior to clinical diagnosis. This approach achieved a sensitivity and specificity of 50% indicating poor performance of selected biomarkers in diagnosis breast cancer.

Santillán-Benítez et al. [9] utilized BMI, leptin, L/A ratio and CA 15-3 all together as reliable biomarkers for breast cancer. They analyzed 88 female patients' serum levels of leptin, adiponectin and CA 15-3 along with anthropometric and biochemical features. All feature values were calculated with a 95% confidence interval. Their results showed that the anthropometric features age ($p \leq 0.001$), weight ($p \leq 0.05$) and waist circumference ($p \leq 0.02$) having comparatively higher values in patients with breast cancer than in patients without it. Further, using the 75th percentile set points for BMI, leptin, L/A ratio and CA 15-3 together can be a reliable approach to predict high risk for developing breast cancer.

The diagnostic value of serum resistin in postmenopausal breast cancer (PBC) cases considering clinico-pathological features, serum tumor markers, anthropometric, metabolic and inflammatory features is examined in Dalamaga et al. [10]. Authors have selected 103 postmenopausal women with incident invasive breast cancer, 103 matched controls and 51 with benign breast lesion. Features like serum resistance, tumor markers, anthropometric, and inflammatory factors were studied. The results indicated that the value of serum resistance was significantly higher in breast cancer group when compared with the other groups. Authors have also found significant correlation among serum resistin with all measured features except with anthropometric, metabolic features and hormone receptor status. Further, analysis using multivariable regression analysis indicated that the strongest determinants of diagnosis were serum IL and cancer stage.

Kloten et al. [11] aimed at identifying a novel biomarker for the screening of blood-based breast cancer by examining the DNA methylation in circulating free DNA (cfDNA) from the blood of breast cancer patients and control subjects. Authors assessed the promoter methylation of seven putative tumor-suppressor genes (SFRP1, SFRP2, SFRP5, ITIH5, WIF1, DKK3, and RASSF1A) which were extracted from cfDNA. Their results demonstrated ITIH5 and DKK3 promoter methylation as potential biomarkers achieving 41% sensitivity with a specificity of 93% and 100% in healthy and benign disease controls, respectively in both test and validation process. It is further observed that the association of these genes with RASSF1A methylation increased the sensitivity to 67%. This study concludes that these three genes (ITIH5 and DKK3 with RASSF1A methylation) could be used as a potential biomarker in screening breast cancer at early stages.

In Zhu et al. [12], a multiple logistic regression model using tumor hemoglobin features measured by ultrasound-guided near-infrared optical tomography (US-NIR) with standard pathologic tumor properties to predict pathologic response of patients before neo adjuvant chemotherapy treatment (NAC) is developed. Thirty-four patient's data were split into 30 groups of training (24 tumors) and testing (12 tumors). Tumor vascularity was assessed via US-NIR measurements of total hemoglobin (tHb), oxygenated hemoglobin (oxyHb) and deoxygenated hemoglobin (deoxyHb) concentrations before starting treatment. Tumor type, Nottingham score, mitotic index, the estrogen and progesterone receptors and human epidermal growth factor receptor 2 feature variables obtained from biopsy were also utilized in modeling. They graded patients' pathologic response on the basis of Miller-Payne system. Their approach achieved average sensitivity of 56.8%, average specificity of 88.9%, average positive predictive value (PPV) of 84.8%, average negative predictive value (NPV) of 70.9% and average area under curve (AUC) of 84.0%, respectively, while using tumor pathologic variables alone for testing data. However, when tHb was included as an additional predictor, the results improved to 79%, 94%, 90%, 86% and 92.4%, respectively, while when oxyHb was included as an additional feature with tumor variables, the results obtained were 77%, 85%, 83%, 83% and 90.6%, respectively. The results of this study indicated that the addition of tHb or oxyHb improved the prediction sensitivity, NPV and AUC when compared with using tumor pathological features alone. However, the model was tested for a small cohort of patients thereby questioning the reliability of the proposed expert system.

Zheng et al. [13] proposed a hybrid approach using *k*-means and support vector machine to recognize the hidden patterns separately for the benign and malignant tumors for their classification. The membership to the hidden patterns of each sample is treated as a new feature for the job of classification. Their approach obtained classification accuracy of 97.38% with 10-fold cross validation when tested on the Wisconsin Diagnostic Breast Cancer (WDBC) data set. Out of 32 original features they extracted six features of high significance to reduce the training time of classifier.

The role of serum irisin for diagnosis of breast cancer is investigated in Provatopoulou et al. [14]. In their study 101 female patients with invasive ductal breast cancer were studied along with 51 healthy women. Logistic regression was used to analyze the affiliation of irisin for breast cancer diagnosis. They found that the serum levels of irisin were significantly lower ($p < 0.001$) in breast cancer patients compared to healthy ones (2.47 ± 0.57 and 3.24 ± 0.66 μg/ml respectively). It was also observed that a 1 unit increase in irisin levels reduces the probability of breast cancer nearly by 90%. The discriminative performance of irisin was obtained at a cut-off point of 3.21 μg/ml with 62.7% sensitivity and 91.1% specificity in addition with strong association with tumor stage and mild associations with tumor size and lymph node metastasis ($p < 0.05$, $p < 0.01$, $p < 0.01$ respectively).

The diagnostic and predicted value of serum levels of leptin, resistin and visfatin in postmenopausal breast cancer is investigated in Assiri and Kamel [15]. The proposed study was conducted on 298 postmenopausal females and grouped into 3 categories: 110 breast cancer (BC), 89 matched healthy controls (HC) and 99 with benign breast lesion (BBL). It was found that the serum levels of leptin, resistin and visfatin were significantly higher in BC groups compared to the other two groups. A survey on existing databases that have been employed for detection of biomarkers for cancer of breast is reported in Lee and Moon [16]. It was concluded that advancements in bioinformatics tools have played significant role in differentiating structures of molecules and genome in cancers along with drug response. The information provided with these databases can be used to comprehend genetic and epigenetic profiles of cancer. A model of genetic algorithm with rotation forest for the accurate diagnosis of breast cancer is proposed in Alickovic´ and Subasi [17]. The performance of the proposed model was also compared with several other data mining techniques. For the performance evaluation, two different datasets of Wisconsin breast cancer data were used. A total of 39 attributes were used along with performance measures such as classification accuracy, area under curve and *F*-measure. The proposed method achieved highest classification accuracy of 99.48% compared to other approaches investigated in this study.

The issue of heterogeneity in patient samples can significantly affect the outcomes of breast cancer prediction. This issue is addressed in Choi et al. [18]. The authors computed principal components of gene expressions followed by clustering using *K*-means algorithm. Weights of edges where then computed for each cluster of samples followed by ranking of genes by using modified page rank. The results of this study indicated that this technique can successfully identify prognostic genes in the networks form that offer significant information regarding molecular functions associated to progression of cancer. A detailed review of prognostic and predictive biomarkers that can help in optimizing decision making of therapy in newly diagnosed breast cancer patients is reported in Nicolini et al. [19]. The review suggested that recognized traditional prognostic factors together with certified prognostic/predictive biomarkers should be utilized for treatment of newly detected breast cancer patients in future.

A prediction model for diagnosis of breast cancer using features collected during usual blood analysis is reported in Patrício et al. [20]. Clinical and anthropometric data like age, BMI, glucose, insulin, HOMA, leptin, adiponectin, resistin and MCP-1 were collected. Machine learning algorithms includes logistic regression, Random forest and support vector machine (SVM) were implemented on the collected dataset and the end results were examined with the help of Monte-Carlo cross validation technique for sensitivity, specificity and area under curve (AUC) measures. A sensitivity ranging between 82% and 88% and specificity ranging between 85% and 90% is obtained when SVM predictive model with glucose, BMI, age and resistin as predictors were used in identifying the presence of breast cancer. Phillips et al. [21] investigated volatile organic compounds of breath as biomarkers of breast cancer indication. Experiments were conducted on 54 women with histopathologically confirmed breast cancer and 124 healthy controls. Their results indicate that volatile organic compounds samples of breath gathered with ultra-clean balloons can be used as reliable biomarkers for breast cancer detection. Single drug biomarker representations for four customary chemotherapy agents: paclitaxel (T), 5-fluorouracil (F), doxorubicin (A) and cyclophosphamide (C) is created to foresee reaction and endurance of ER− breast cancer patients treated with combination chemotherapies in Chen et al. [22]. Background parenchymal uptake (BPU) with molecular breast imaging (MBI) as a potential biomarker for breast cancer risk prediction is discussed in Besutti et al. [23]. The authors concluded that MBI BPU can append new data to the risk factors of large panorama of breast cancer to be utilized for customized screening. Examining static single-biopsy next-generation sequencing and spatiotemporal discovery of genomic clones in tumoral and liquid biopsies and blood related markers for early detection of breast cancer is looked out as an important progress [24]. Recognition of novel therapeutic targets for triple-negative breast cancer (TNBC) is reported in Feng et al. [25]. Their research revealed the expression profiles of tRNA-derived small non coding RNAs (tDRs) in TNBC and non-TNBC cancer stem cells. These expression profiles can provide useful information to know the tDR-000620 expression which is accountable for the destructive phenotype of breast cancer stem cells. A review of explicit roles of biomarkers related to imaging of breast in research is conducted in Weaver and Leung [26]. The review concluded that amalgamation of breast imaging with associated biomedical area and the availability of large open source databases of clinical, molecular, and imaging biomarkers in future is needed for guiding breast cancer research.

Having reviewed the aforementioned studies, it is found that different types of biomarkers have been identified for breast cancer detection such as those from molecular imaging, volatile organic compounds of breath, blood tests, gene expressions etc. Few studies also suggested that blood related biomarkers are looked out as an important development for early detection of breast cancer. Further, it is concluded that data mining and machine learning techniques such as SVM, random forest, principal component analysis, logistic regression etc. have been employed by for analyzing clinical data, anthropometric measures, image measures pertaining to breast cancer. However, there is no systematic study signifying the role of feature selection to identify reliable clinical and anthropometric measures. Some of the reported studies utilize PCA for eliminating redundant and unimportant features. However, uses of lone assessment measure have demonstrated restricted performance in CAD systems. Thus, a systematic study and evaluation of different feature selection techniques in identifying relevant clinical and anthropometric measures is required. Some studies like [20] did not evaluate their database for different training and test data division protocols. The hold out data division protocol used in some of these is not a preferred approach to evaluate classifier models because there is a probability that some samples may never be the part of training or testing group due to random selection of training and test samples. Hence, classifier model must be evaluated for different data division protocols. To overcome these limitations, this paper implements and evaluates various traditional feature selection techniques to identify reliable clinical and anthropometric biomarkers for breast cancer prediction using kernel based SVM,Naïve Bayesian, linear discriminant, quadratic discriminant,logistic regression, K-nearest neighbor (K-NN) and random forest. The proposed model is evaluated under different data division protocols and kernel properties.

1.2. *Contributions and organization of the paper*

The contributions of this paper are summarized as follows:

(i) A recent database of clinical and anthropometric measurements acquired from normal individuals and individuals suffering from breast cancer [20] is utilized and evaluated in machine learning framework.

(ii) Extensive experimental investigation using statistical significance analysis techniques such as quantile–quantile probability plots, Mann–Whitney *U*-test, independent samples *t*-test is conducted to estimate the significance of clinical and anthropometric biomarkers.

(iii) Implementation and evaluation of various machine learning components such as feature selection, data division strategies and classification to identify apposite biomarkers for breast cancer prediction using clinical and anthropometric features is conducted.

The rest of the paper is organized as follows. Section 2 presents material and methods used in this study. Section 3 presents results and discussion followed by conclusion and future directions in Section 4.

2. Material and methods

The proposed archetype for prediction of presence of breast cancer using anthropometric and clinical features in machine learning paradigm is shown in Fig. 1. It consists of two phases,

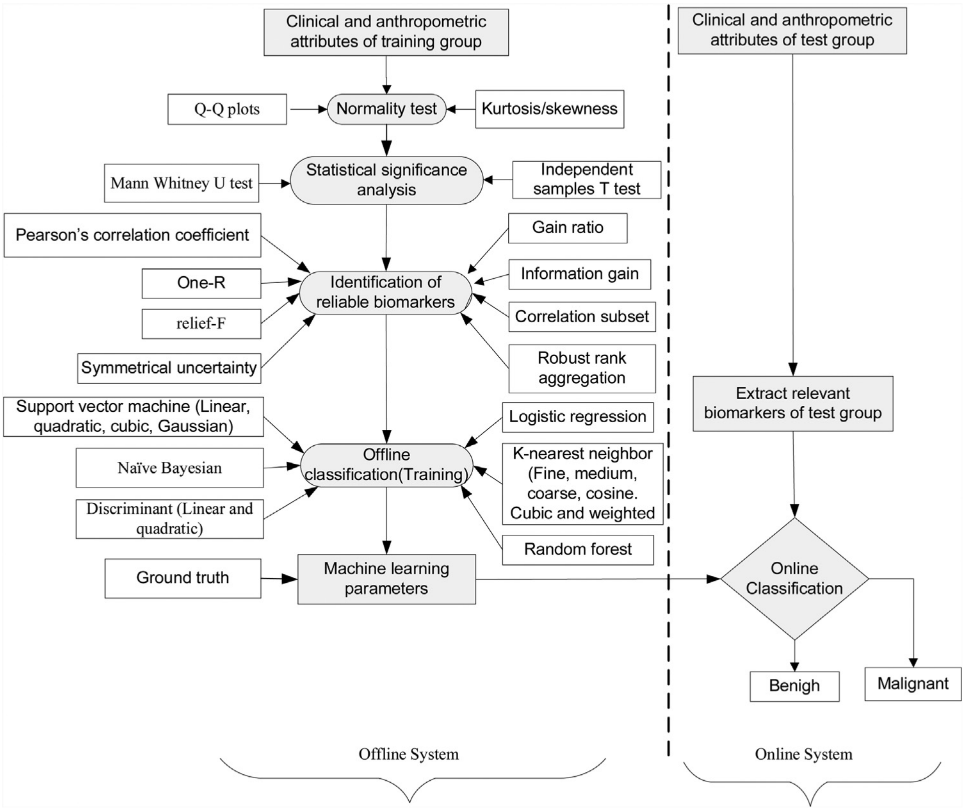

Clinical and anthropometric attributes of training group
Clinical and anthropometric attributes of test group
Q-Q plots
Normality test
Kurtosis/skewness
Mann Whitney U test
Statistical significance analysis
Independent samples T test
Pearson's correlation coefficient
Gain ratio
One-R
Identification of reliable biomarkers
Information gain
relief-F
Correlation subset
Symmetrical uncertainty
Robust rank aggregation
Extract relevant biomarkers of test group
Support vector machine (Linear, quadratic, cubic, Gaussian)
Logistic regression
Naïve Bayesian
Offline classification(Training)
K-nearest neighbor (Fine, medium, coarse, cosine. Cubic and weighted
Discriminant (Linear and quadratic)
Random forest
Ground truth
Machine learning parameters
Online Classification
Benigh
Malignant
Offline System
Online System

Fig. 1 – Proposed strategy for prediction of breast cancer using anthropometric and clinical features.

alienated by the dotted line. The left side of this figure illustrates an offline scheme which includes quantitative analysis, statistical significance analysis, feature selection and building a classifier model using ground truth. On the other hand, the right side represents an online mode that predicts the class label of test cases based on the offline training variables. The offline system (training phase) usually involves training of machine learning model using features extracted. In offline mode, both the features and its category/class (ground truth) are provided to train the classifier model in supervised manner. In online system (testing phase), the machine learning model developed in offline mode is tested/evaluated on new cases. Each of these steps is explained in detail in the following sections along with details of dataset used in this study.

2.1. Data

The database used in this study consists of several clinical and anthropometric measures such as age, body mass index (BMI), glucose, insulin, HOMA, leptin, adiponectin, resistin and MCP.1 obtained from 52 healthy and 64 suffering from breast cancer. The diagnosis was confirmed using mammography followed by biopsy. The details of the database can be found in Patrício et al. [20].

2.2. Descriptive analysis

To start with the analysis, initially various clinical and anthropometric measures are examined by obtaining their summary statistics. This includes statistical measures such as mean standard error (mean), median, variance, standard deviation, range, skewness and kurtosis. Further, error plots such as Bland–Altman plots and regression plots were also used to analyze the variability of different measurements among both groups. The main aim of descriptive analysis is to identify atypical observations that may create problems in later analysis of the data.

2.3. Normality test

Normality test is conducted to verify whether the data has been drawn from a normally distributed population. Several statistical tests entail a sample population which is normally distributed. Among various available methods, quantile–quantile probability plot (Q–Q plot) is utilized to assess the normality of the data. It is a plot between quantiles estimated from standard normal distribution against the observed quantiles. If the data points corresponding to a particular group lie approximately on the reference line, the data is said to have normal distribution. Data points above the reference line specify that the observed quantiles are lesser than estimated and vice versa [27].

2.4. Statistical significance analysis

Statistical significance of various clinical and anthropometric features is assessed using two methods namely Independent samples *t*-test and Mann–Whitney *U*-test. Independent samples *t*-test is used to test the null hypothesis that the mean of features of controlled and diseased groups are the same. However, this test assumes data to be normally distributed. Thus, Mann–Whitney *U*-test which is useful for non-uniformly distributed data is also conducted.

2.5. Determining relevant biomarkers using feature selection techniques

Selection of reliable biomarkers plays crucial role in breast cancer representation and classification using machine learning techniques. Further, it is very vital step in machine learning paradigm, since; poor selection of features (biomarkers) can lead to an inaccurate prediction model. Feature selection is a procedure of choosing the most pertinent features having high discriminatory power in differentiating cancer types. Broadly, feature selection techniques are classified in to two types namely, filter method and wrapper methods.

Filter methods determines rank of features by using some principal criterion [28]. In this study, some popular ranking criteria such as Pearson's correlation coefficient (P) [28], gain ratio (GR) [29], information gain (IG) [30], 1R [31], Relief-F (RF) [32] and symmetrical uncertainty (SU) [33] have been utilized and evaluated for selection of relevant biomarkers.

Wrapper methods initially generate various possible subsets (say, 2^n) from given number of features (say, n) and then evaluate them using a specific objective function. The subset of features with top performance is kept while all other subsets are discarded. Compared to filter method, wrapper method suffers from high computational complexity, particularly if n is very large. In this study, a correlation based wrapper feature selection (CFS) [34] approach is used and evaluated for selection of most reliable subset of biomarkers. Further, the comparative evaluation of various feature selection techniques based on filter and wrapper approach is conducted. Further, a hybrid approach called as robust rank aggregation (RRA) is also implemented and evaluated [35]

2.6. Classification

The final phase of any CAD system is the classifier which maps input feature vectors $\boldsymbol{x} \in \boldsymbol{X}$ to output class labels $\boldsymbol{y} \in \{1, \ldots, \boldsymbol{n}\}$, where $\boldsymbol{X}$ is the feature space and $\boldsymbol{n}$ is the total number of classes. Classification techniques are broadly classified into two types namely, supervised and unsupervised. In supervised classifier, the training samples are supplied along with its class labels. The class label of unknown cases i.e. the test samples is then determined based on the parameters of trained classifier model. In this study, some of the most popular supervised classifiers such as kernel based support vector machine (SVM), Naïve Bayesian, linear discriminant, quadratic discriminant, logistic regression, *K*-nearest neighbor (*K*-NN) and random forest are used. Support vector machine is the state-of-the-art approach widely used for classification and regression problems. It separates two classes of a sample by constructing a hyperplane to distinguish class members from non-members [36]. Since most of the real world problems are nonlinear, to build non-linear classifiers for classifying the data points, kernel based SVM is used [37]. The input data space is mapped into higher dimension feature space denoted by $\varphi : X \rightarrow \varphi(X)$. The kernel function is the inner product of the data variables in feature space such that $k(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j)$. Various kernel used in this study are shown in Table 1.

Naïve Bayes' classifier is a statistical approach based on Bayes' theorem. In this technique, the probability that a particular sample belongs to a particular class is determined by assuming that each value of attribute independently contributes to the specified class in supervised manner. The third classifier which is implemented and evaluated in this study is discriminant analysis. Originally introduced by R. Fisher, this technique has been successfully applied to solve many classification problems [38]. In this approach, the Gaussian distribution associated parameters of every category are predicted using some fitting function and a priori probabilities of belongingness to class. It is of two type's namely linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). The main objective in LDA is to determine a linear transformation that best distinguish amongst classes followed by classification in the transformed space based on some predefined metric such as Euclidean distance [39]. In LDA, linear combination of predictor variables are used to model the classifier unlike QDA in which non-linear combination of predictor variables are used for building the classifier model.

The fourth classifier that is evaluated in this study is logistic regression. In this classification technique a threshold level is selected on the basis of which the class corresponding to set of predictors is determined. The predictions are transformed using the logistic function. The fifth classifier implemented for breast cancer prediction is K-nearest neighbor (K-NN). K-nearest neighbor performs classification task using some distance metric from nearest sample (K is number of neighbors). In this work, the value of k is set to $k = 1$, 10, and 100 with different distance measures such as Euclidean, cosine, cubic, and weighted for binary classification. This results in different K-NN types based on number of neighbors and distance measure used namely, fine K-NN ($K = 1$, distance metric: Euclidean), medium K-NN ($K = 10$, distance metric: Euclidean), coarse K-NN ($K = 100$, distance metric: Euclidean), cosine K-NN ($K = 10$, distance metric: cosine), cubic K-NN ($K = 10$, distance metric: Minkowski (cubic)) and weighted K-NN ($K = 10$, distance metric: Euclidean with squared inverse distance weight). The last classifier which is evaluated in this study is random forest. Random forests are classifiers based on ensemble learning technique that work by constructing a large number of decision trees. To construct the tree, each node is divided using the best among the set of extracted features randomly chosen at that node. For classification of a new sample, the classifier takes the feature vector at input, classifies it with each tree in the forest, and determines the class label based on majority voting [40].

2.7. *Performance evaluation*

2.7.1. *Performance measures*

The various performance metrics used to evaluate the classifiers are classification accuracy, sensitivity and specificity, as shown in Table 2. The symbols η_{tp}, η_{fp}, η_{fn} and η_{tn} are number of true positives, false positives, false negatives and true negatives classified by classifier, respectively [5]. Along with these performance measures, area under receiver operating characteristics (AUC) is also used to compare classifier models. It is defined as area of the curve obtained by plotting the true positive rate (TPR) i.e. sensitivity versus the false positive rate (FPR) at various thresholds.

2.7.2. *Data division protocol*

To evaluate the performance of proposed classifier model the whole dataset is divided into two parts, one for training the classifier and other for testing it. Two different data division namely, holdout and k-fold cross validation are used in this study. In holdout protocol, the dataset is divided into two groups in random manner. One of these groups is used for training while other group is used for testing the classifier model. In this study, 67% of the samples were used for training while 33% of the samples were used for evaluating the classifier model. The second data division protocol i.e. k-fold cross validation is the most popular and extensively acknowledged by research community. In this approach, the whole dataset was divided in to 'k' groups, consisting of approximately equal number of samples. Out of 'k' groups, '$k - 1$' groups are used for training the classifier model while remaining one group is used for testing [5]. The process is repeated 'k' times and average performance over 'k' rounds is calculated. In this study, experiments were conducted with value of $k = 5$ and 10. Further, for both hold out and k-fold, the whole process of training and testing is repeated 5 times and average performance is calculated.

3. Results and discussions

Finding relevant biomarkers from routine clinical attributes is significant for early detection of breast cancer, particularly in

Table 1 – Classification techniques and their parameters used in this study.

Type of classification method	Kernel type	Description
Linear SVM	Linear kernel	$k(x_i, x_j) = (x_i \cdot x_j)$
Quadratic SVM	Polynomial kernel	$k(x_i, x_j) = (1 + x_i \cdot x_j)^2$
Cubic SVM	Polynomial kernel	$k(x_i, x_j) = (1 + x_i \cdot x_j)^3$
Fine Gaussian SVM	Gaussian radial basis function	$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), \quad \sigma = 0.75$
Medium Gaussian SVM	Gaussian radial basis function	$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), \quad \sigma = 3$
Course Gaussian SVM	Gaussian radial basis function	$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), \quad \sigma = 12$

Table 2 – Performance measures used to evaluate classifier model.

Performance measures (%)	Definition
Overall classification accuracy (A)	$\frac{\eta_{tp}+\eta_{tn}}{\eta_{tp}+\eta_{fn}+\eta_{tn}+\eta_{fp}}\times 100$
Sensitivity	$\frac{\eta_{tp}}{\eta_{fn}+\eta_{tp}}\times 100$
Specificity	$\frac{\eta_{tn}}{\eta_{tn}+\eta_{fp}}\times 100$

low resource settings. There is an urgent need of approaches for detection of function-associated unswerving biomarkers which can be used to evaluate treatment response and disease progression. Several studies specify that blood and other body fluids can be useful in identification of clinically helpful biomarkers for breast cancer detection and control [41]. Machine learning techniques for determining such biomarkers have thus gained much attention in recent years. Besides its huge impact in the medical industry, there are growing numbers of multinational companies becoming active in this research field. However, the existing studies suffer from several drawbacks due to precincts posed by poor quality and incomplete record of medical data, large size of data and regulatory and commercial standards restraining their application in routine clinical practice. Thus, enhancing the performance of such systems has turned out to be a crucial research task.

The vital elements of computerized disease diagnosis systems in machine learning paradigm includes: (a) the choice of appropriate feature selection technique to determine significant biomarkers to discriminate diseased and control group; and (b) the choice of appropriate classifier model for disease prediction at early stages. This article investigates the efficacy of clinical features recorded during routine blood analysis for predicting breast cancer using predictive machine learning techniques. Further, role of various machine learning modules such as biomarker selection, data division protocols and classification is studied in detail.

3.1. *Results of descriptive analysis*

Table 3 shows the descriptive statistics of various clinical attributes obtained from participants consisting of two groups i.e. group 1 (controlled) and group 2 (diseased) obtained using statistical package for the social sciences (SPSS) software. It is found that features like glucose, insulin, HOMA, resistin and MCP.1 have good variability among two groups. This is evident from values of mean, standard error, median, standard deviation, range, skewness and kurtosis. On the other hand, features like age, BMI, leptin and adiponectin show least variability among the two groups.

Figs. 2 and 3 show regression and Bland Altman plots respectively for various measures in the database using MedCalc software for 95% confidence interval. Regression plot displays scatter plot of variables along with regression line (blue colored line), 95% confidence interval and 95% prediction interval.

Regression plots are useful in understanding how one variable changes with another. Further, it is used to determine how one variable is likely to be when the other variable is known within the limits of the scatter diagram. Bland–Altman plot indicates the bias or average difference between the variables among two groups. Figs. 2 and 3 reveal high degree of disagreement between variables glucose, insulin, HOMA and resistin between two groups. Thus, these features may be useful in discriminating benign and malignant breast cancer in women.

3.2. *Results of normality analysis*

Fig. 4 shows Q–Q plot of various features obtained using SPSS software. It is observed that for most of the features some degree of positive skewness is detected except age (group 1) which is negatively skewed. This is evident from most of the Q–Q plots that small and very large quantiles are higher than estimated herewith being more prominent for the lower quantiles. Similar inferences can also be made from values

Table 3 – Descriptive statistics of data.

Features	Statistics								
	Group	Mean	Standard error (mean)	Median	Variance	Standard deviation	Range	Skewness	Kurtosis
Age	1	58.08	2.629	65	359.406	18.958	65	−0.276	−1.259
	2	56.67	1.687	53.00	182.065	13.493	52	0.532	−0.757
BMI	1	28.317	0.752	27.694	29.457	54.274	19.908	0.152	−1.187
	2	26.984	0.577	27.408	21.348	4.620	18.739	0.056	−0.832
Glucose	1	88.23	1.413	87	103.867	10.192	58	0.302	1.242
	2	105.56	3.320	98.50	705.298	26.557	131	2.157	5.319
Insulin	1	69.337	0.674	5.483	23.618	4.859	23.504	2.414	6.178
	2	12.513	1.539	7.580	151.727	12.317	56.028	1.960	3.773
HOMA	1	1.552	0. 168	1.139	1.484	1.218	6.644	2.693	8.730
	2	3.623	0.573	2.052	21.058	4.588	24.542	2.911	9.669
Leptin	1	26.637	2.681	21.494	373.831	19.334	79.171	1.152	0.794
	2	26.596	24.015	18.877	369.118	19.212	83.946	1.471	2.173
Adiponectin	1	10.328	1.058	8.127	58.236	7.631	35.845	2.089	4.617
	2	10.061	0.773	8.446	38.309	6.189	32.094	1.373	2.330
Resistin	1	11.614	1.587	8.929	131.035	11.447	78.808	4.796	28.688
	2	17.253	1.579	14.372	159.693	12.636	52.005	1.526	2.153
MCP.1	1	499.730	40.526	471.322	85,410	292.242	1210.2	0.742	0.094
	2	563.016	48	465.37	147,500	384.001	1608.40	1.569	2.632

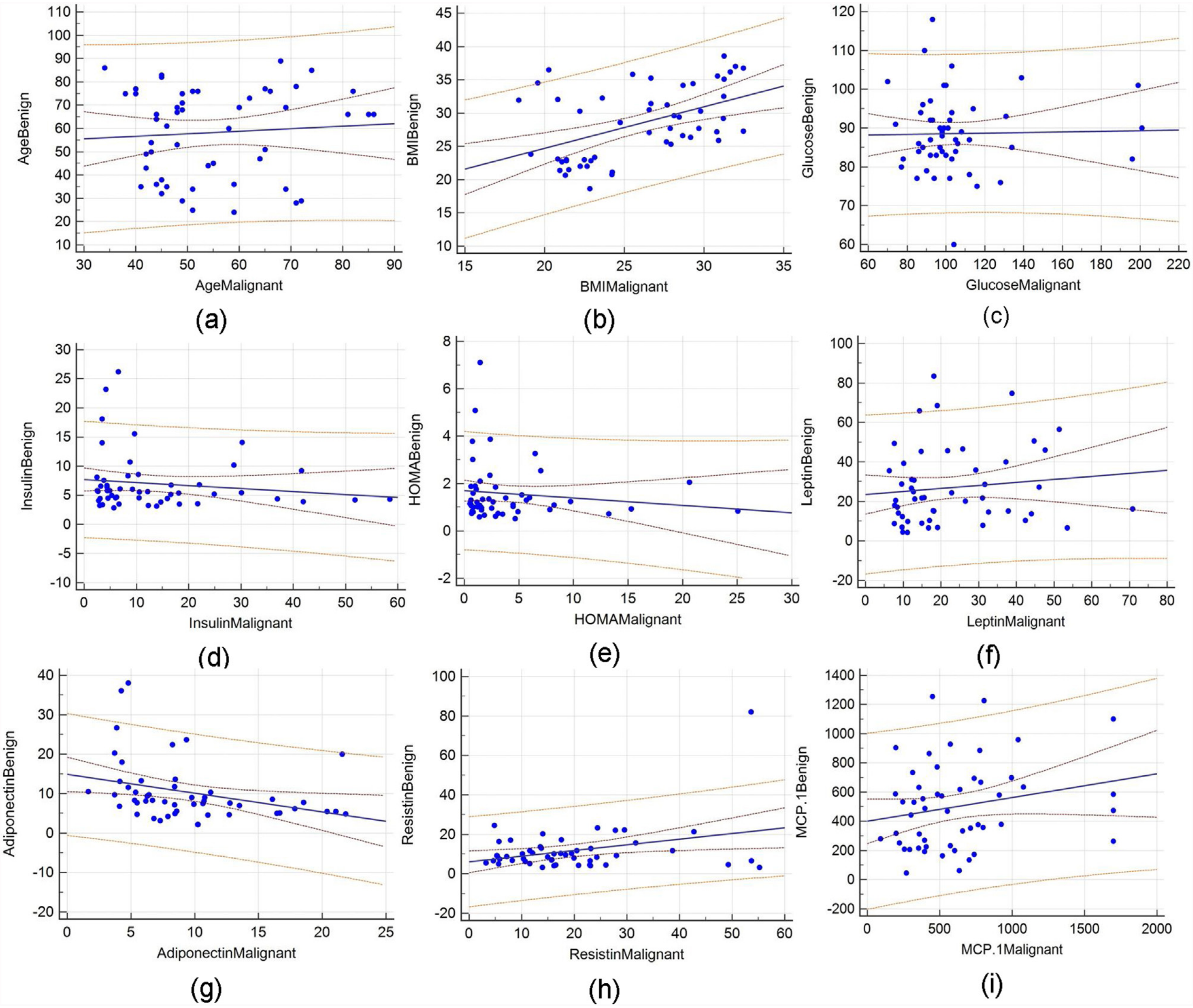


Fig. 2 – Regression plots of various anthropometric and clinical features among two groups.

of skewness and kurtosis in Table 3. It is thus concluded that normality assumptions are not met by the dataset used in this study.

3.3. *Results of statistical significance analysis*

Two popular techniques namely independent samples *t*-test and Mann–Whitney *U*-test were used evaluate statistical significance of various clinical and anthropometric variables. These tests were conducted using SPSS software for 95% confidence interval. All variables with *p*-values less than 0.05 were considered statistically significant.

Independent samples *t*-test is used to test the null hypothesis that the mean of features of controlled and diseased groups are the same. For instance, to conduct independent samples *t*-test the results of normality test as discussed in previous section is neglected. Table 4 shows the results of two versions of *t*-test: (i) assuming equal variances within two groups (first row), and (ii) assuming unequal variances within two groups (second row). It is observed that, no significant statistical difference exists for features such as age, BMI, leptin, adiponectin and MCP.1 since *p*-value is greater than 0.05 for all these features. On the other hand, features like glucose, insulin, HOMA and resistin are found to be statistically significant with *p*-value <0.05 for all these features. The results of *t*-test also suggest that there is a considerable difference in the glucose levels of two groups (*p*-value <0.001).

The results of normality test using Q–Q plots indicated that all the clinical features considered in this study do not follow normal distribution. Thus, it is more useful to conduct non parametric test like Mann–Whitney *U*-test which does not rely on assumptions of normal distribution unlike independent samples *t*-test. The corresponding results are shown in third row of Table 4. The results of Mann–Whitney *U*-test are found to be consistent with that of independent samples *t*-test i.e. clinical features such as glucose, insulin, HOMA and resistin are found to statistically significant with *p*-value <0.05. Based on the results of statistically significant analysis, features like

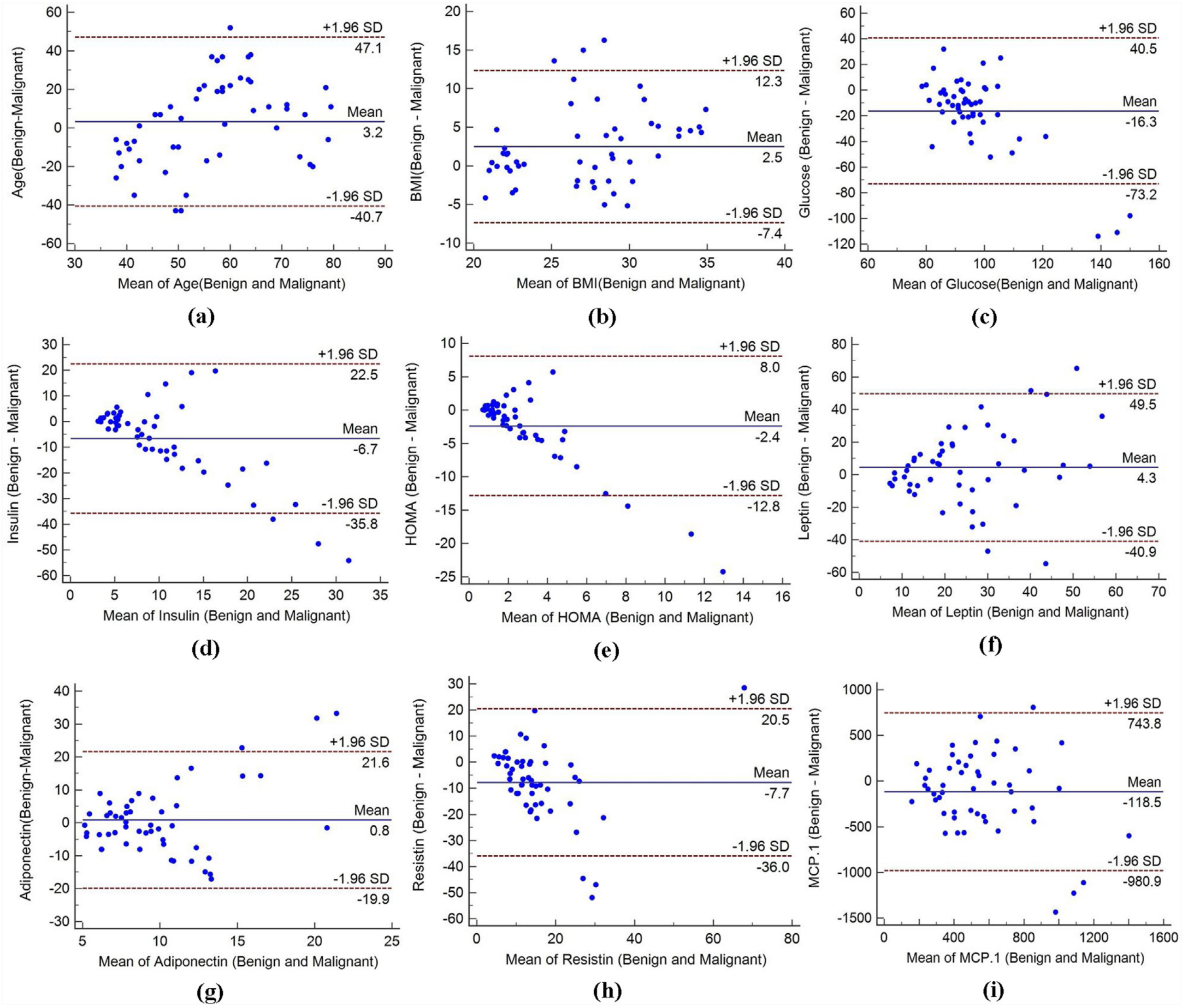


Fig. 3 – Bland–Altman plots of various anthropometric and clinical features among two groups.

glucose, insulin, HOMA and resistin are identified as reliable biomarkers for breast cancer risk prediction.

3.4. *Results of feature selection*

Table 5 shows the results of various feature selection techniques. In filter based methods, features are arranged in decreasing order of their rank while in wrapper based method, best subset of features is selected. It is found that rank assigned to various features by different feature selection techniques is slightly different. For example, if P is used as principal criteria, glucose is considered as most reliable biomarker. On the other hand, if RF is used as principal criteria, age is considered as most reliable biomarker. Similarly, HOMA is assigned rank 2 if P is used as principal criteria while it is assigned rank 5 if GR or IG is used as principal criteria. It is thus concluded that relying on one principal criterion may not always result in optimal subset of biomarkers. An optimal subset of biomarkers elected using one assessment measure may not be the similar as that using other. The only way to assure high accuracy is to examine the classifier on various feature subsets, obtained from different raking measures [42]. The performance of various feature selection techniques are evaluated using kernel based SVM. The corresponding results are presented and discussed in the forthcoming section.

3.5. *Results of classification using kernel based SVM*

This section presents the results of different SVM classifiers with and without using feature selection step. Three performance measures (accuracy, sensitivity and specificity) are used for evaluation under three data division protocols namely hold out, 5-fold and 10-fold. This step is implemented using MATLAB® software. Table 6 shows the performance of different classifiers without using feature selection technique i.e. all the nine clinical and anthropometric features are supplied to the input of classifier. It is found that medium

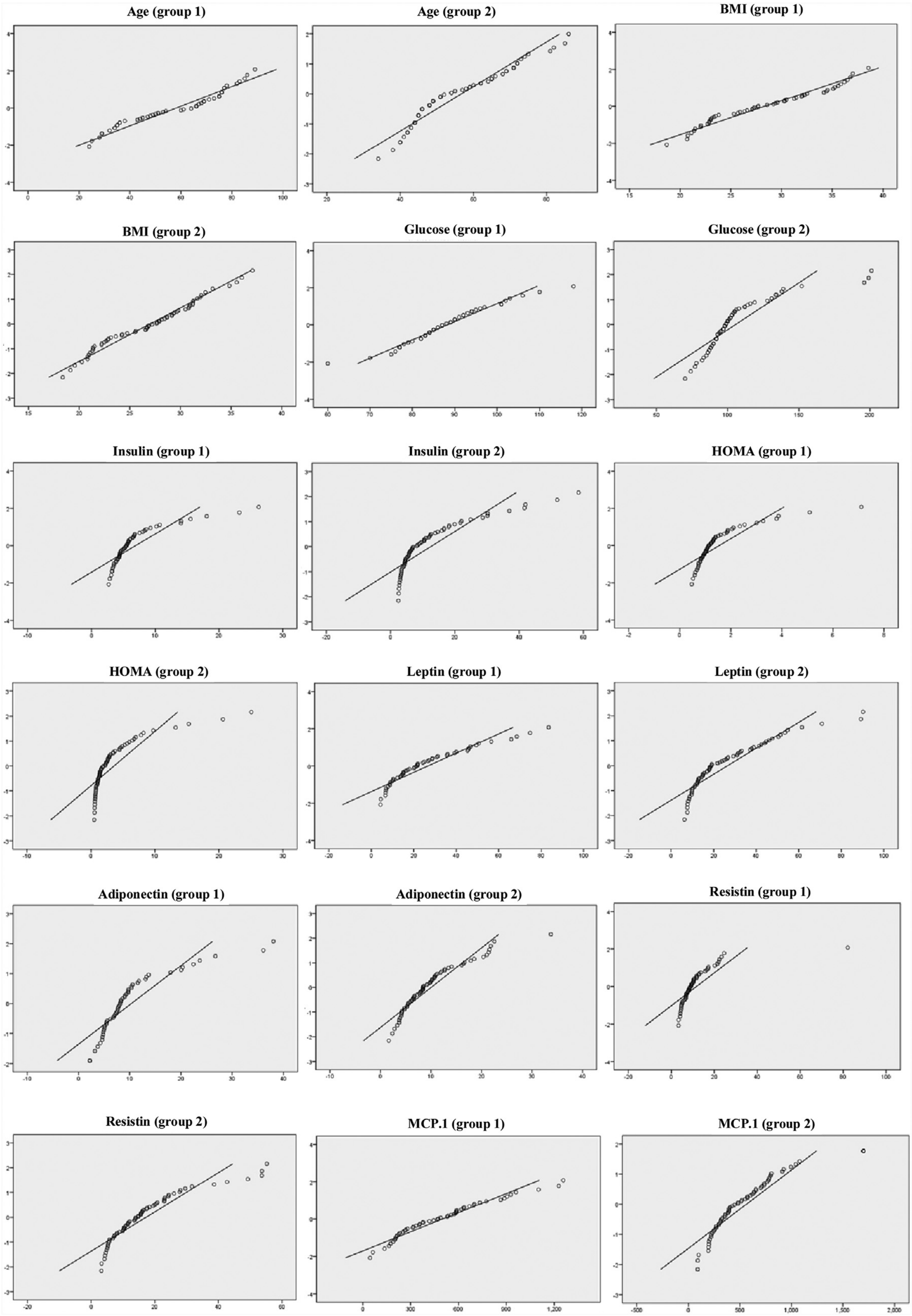


Fig. 4 – Q–Q plots for various features. *x*-axis represents the observed value and *y*-axis represents the expected normal value.

Table 4 – Results of independent samples *t*-test and Mann–Whitney *U*-test (95% confidence interval).

p-value	Features								
	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
p-value[a]	0.642	0.156	0.000	0.003	0.002	0.991	0.835	0.014	0.329
p-value[b]	0.654	0.163	0.000	0.001	0.001	0.991	0.839	0.013	0.316
p-value[c]	0.477	0.201	0.000	0.026	0.003	0.947	0.764	0.002	0.502

The *p*-values shown in the table are obtained using:
[a] independent samples *t*-test with equal variance assumed.
[b] independent samples *t*-test with equal variance not assumed.
[c] Mann–Whitney *U*-test.

Gaussian SVM outperform other classifiers achieving highest classification accuracy of 74.665%, 72.931% and 68.421% under 5-fold, 10-fold and hold out data division protocol respectively. On contrary, course Gaussian SVM perform worst achieving lowest classification accuracy of 56.552%, 56.034% and 55.789% under 5-fold, 10-fold and hold out data division protocols respectively. Thus, it is concluded that highest classification accuracy of 74.655% is achieved without using feature selection.

Table 7 shows the performance of different classifiers when top 3 features namely, glucose, HOMA and Insulin selected by Pearson's correlation coefficient (P) is supplied as input to the classifier. It is found that course Gaussian SVM classifier under hold out data division protocol outperform others achieving classification accuracy of 72.105%. However, under 5-fold and 10-fold cross validation schemes, linear SVM and medium Gaussian SVM achieve higher classification accuracy of 70.345% and 70.172% respectively. The worst performance is demonstrated by cubic SVM displaying lowest classification accuracy under all data division schemes.

Table 8 shows the performance of different classifiers when top 3 features namely, glucose, age and insulin selected by Gain ratio(GR), Information gain (IG) and Symmetrical uncertainty (SU) is supplied as input to the classifier. It is found that medium Gaussian SVM classifier outperform others under all data division schemes. It achieves classification accuracy of 73.966%, 73.793% and 76.316% under 5-fold, 10-fold and hold out data division protocols respectively. On the other hand, categories of test samples predicted by cubic SVM match least with ground truth categories resulting in its lowest classification accuracy under all three data division schemes.

Table 9 shows the performance of different classifiers when top 3 features namely, glucose, age and resistin selected by Gain by1R and Relief-F (RF) is supplied as input to the classifier. It is found that medium Gaussian SVM classifier outperform others under all data division schemes. It achieves highest classification accuracy of 81.897%, 82.388% and 83.684% under 5-fold, 10-fold and hold out data division protocols respectively. On the contrary, coarse Gaussian SVM results in lowest classification accuracy under all three data division schemes.

It is interesting to note here that the feature age as an individual feature was not found statistically significant when analyzed using independent samples test and Mann–Whitney *U*-test. However, when this feature is combined with glucose and resistin (as in present case), the performance of classifier achieves highest values. Further, compared to all other feature combinations, the combination of glucose, age and resistin achieves highest classification accuracy of 83.684%. Table 10 shows the performance of different classifiers when best subset of features selected by correlation based wrapper feature selection (CFS) and top 3 features selected by Robust rank aggregation (RRA) is supplied as input to the classifier. The feature combination evaluated in this case is age, glucose and HOMA. As in most of the previous cases, it is found that medium Gaussian SVM classifier outperform others under all data division schemes. It achieves highest classification accuracy of 74.655%, 73.621% and 75.263% under 5-fold, 10-fold and hold out data division protocols respectively.

From the results of Tables 6–10, it is concluded that the feature combination age, glucose and resistin achieve high classification accuracy. To study and confirm the impact of these biomarkers on breast cancer prediction, some other popular classifiers such as Naïve Bayesian, linear

Table 5 – Results of various feature selection techniques.

Name of feature selection technique	Category of feature selection method	Selected subset (wrapper method)/selected features in decreasing order of their rank (filter method)
P	Filter	Glucose, HOMA, insulin, resistin, BMI, MCP.1, age, adiponectin, leptin
GR	Filter	Glucose, age, insulin, BMI, HOMA, resistin, MCP.1, leptin, adiponectin
IG	Filter	Glucose, age, insulin, BMI, HOMA, resistin, MCP.1, leptin, adiponectin
1R	Filter	Glucose, age, resistin, HOMA, insulin, BMI, adiponectin, leptin, MCP.1
RF	Filter	Age, glucose, resistin, BMI, insulin, HOMA, adiponectin, leptin, MCP.1
SU	Filter	Glucose, age, insulin, BMI, HOMA, resistin, MCP.1, leptin, adiponectin
CFS	Wrapper	Age, glucose, HOMA
RRA	–	Glucose, age, HOMA, insulin, resistin, BMI, MCP.1, adiponectin, leptin

Table 6 – Performance of various SVM based classifiers without using feature selection under different data division protocols.

Data division protocol	Classification technique	Performance measures (average value of 5 iterations)			
		Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
5-fold	Linear SVM	71.897	68.438	76.154	0.719
	Quadratic SVM	72.069	73.750	70.000	0.721
	Cubic SVM	71.552	73.438	69.231	0.716
	Fine Gaussian SVM	61.379	94.375	20.769	0.614
	Medium Gaussian SVM	**74.655**	**76.563**	**72.308**	**0.747**
	Course Gaussian SVM	56.552	98.125	5.385	0.566
10-fold	Linear SVM	72.931	67.188	80.000	0.729
	Quadratic SVM	70.000	71.563	68.077	0.700
	Cubic SVM	71.034	73.125	68.462	0.710
	Fine Gaussian SVM	62.069	94.375	22.308	0.621
	Medium Gaussian SVM	**72.931**	**74.375**	**71.154**	0.729
	Course Gaussian SVM	56.034	93.438	10.000	0.560
Hold-out	Linear SVM	66.316	70.476	61.176	0.663
	Quadratic SVM	66.842	69.524	63.529	0.668
	Cubic SVM	62.632	61.905	63.529	0.626
	Fine Gaussian SVM	62.632	96.190	21.176	0.626
	Medium Gaussian SVM	**68.421**	**74.286**	**61.176**	**0.684**
	Course Gaussian SVM	55.789	99.048	2.353	0.558

discriminant, quadratic discriminant, logistic regression, K-nearest neighbor (K-NN) and random forest are also evaluated as discussed in Section 2.5. Table 11 shows the corresponding results. It is observed that when age, glucose and resistin where used as features, the medium K-NN classifier achieves highest classification accuracy of 92.105% under hold out data division protocol. This shows that age, glucose and resistin together can have a significant impact on prediction of breast cancer using machine learning techniques. Other classifiers such as weighted K-NN and cubic-KNN also performed satisfactorily achieving classification accuracy of 81.034% and 83.621% under 5-fold and 10-fold data division protocol respectively. These results are very much comparable to those by SVM. To establish the statistical significance of improvement in classifier performance from 74.665% using medium Gaussian SVM-5-fold (see Table 6) to 92.105% using medium K-NN-hold out (see Table 11), z-statistic is calculated at 95% confidence interval using approach explained in Isaac [43] for test concerning two proportions. The z-statistic is found to be -3.547 with p-value of less that 0.05 at 95% confidence interval. This confirms that the improvement in classification accuracy of medium K-NN classifier over medium Gaussian SVM classifier is statistically significant.

Table 7 – Performance of various SVM based classifiers using top 3 features selected by Pearson's correlation coefficient (P) feature selection under different data division protocols.

Data division protocol	Classification technique	Performance measures (average value of 5 iterations)			
		Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
5-fold	**Linear SVM**	**70.345**	**70.625**	**70.000**	**0.740**
	Quadratic SVM	65.690	62.813	69.231	0.690
	Cubic SVM	56.379	61.875	49.615	0.570
	Fine Gaussian SVM	70.000	83.438	53.462	0.670
	Medium Gaussian SVM	70.172	69.375	71.154	0.750
	Course Gaussian SVM	65.172	65.313	65.000	0.720
10-fold	Linear SVM	69.310	69.688	68.846	0.760
	Quadratic SVM	65.862	63.750	68.462	0.710
	Cubic SVM	49.828	55.313	43.077	0.480
	Fine Gaussian SVM	69.828	83.438	53.077	0.660
	Medium Gaussian SVM	**70.172**	**70.000**	**70.385**	**0.750**
	Course Gaussian SVM	66.724	62.500	71.923	0.740
Hold-out	Linear SVM	70.526	65.714	76.471	0.800
	Quadratic SVM	64.737	59.048	71.765	0.730
	Cubic SVM	54.211	53.333	55.294	0.550
	Fine Gaussian SVM	67.895	79.048	54.118	0.690
	Medium Gaussian SVM	70.882	63.810	79.853	0.790
	Course Gaussian SVM	**72.105**	**78.095**	**64.706**	**0.770**

Table 8 – Performance of various SVM based classifiers using top 3 features selected by gain ratio (GR), information gain (IG) and symmetrical uncertainty (SU) feature selection under different data division protocols.

Data division protocol	Classification technique	Performance measures (average value of 5 iterations)			
		Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
5-fold	Linear SVM	73.448	73.438	73.462	0.766
	Quadratic SVM	71.034	78.125	62.308	0.760
	Cubic SVM	64.138	69.688	57.308	0.690
	Fine Gaussian SVM	70.517	82.813	55.385	0.680
	Medium Gaussian SVM	**73.966**	**83.438**	**62.308**	**0.780**
	Course Gaussian SVM	64.138	82.813	41.154	0.720
10-fold	Linear SVM	71.552	73.438	69.231	0.760
	Quadratic SVM	69.655	79.063	58.077	0.750
	Cubic SVM	63.966	68.438	58.462	0.680
	Fine Gaussian SVM	69.828	82.500	54.231	0.690
	Medium Gaussian SVM	**73.793**	**83.750**	**61.538**	**0.790**
	Course Gaussian SVM	64.310	82.188	42.308	0.730
Hold-out	Linear SVM	67.368	67.619	67.059	0.790
	Quadratic SVM	69.474	74.286	63.529	0.768
	Cubic SVM	57.368	57.143	57.647	0.584
	Fine Gaussian SVM	71.579	86.667	52.941	0.730
	Medium Gaussian SVM	**76.316**	**79.048**	**72.941**	**0.814**
	Course Gaussian SVM	62.632	82.857	37.647	0.750

Table 9 – Performance of various SVM based classifiers using top 3 features selected by1R and Relief-F (RF) feature selection under different data division protocols.

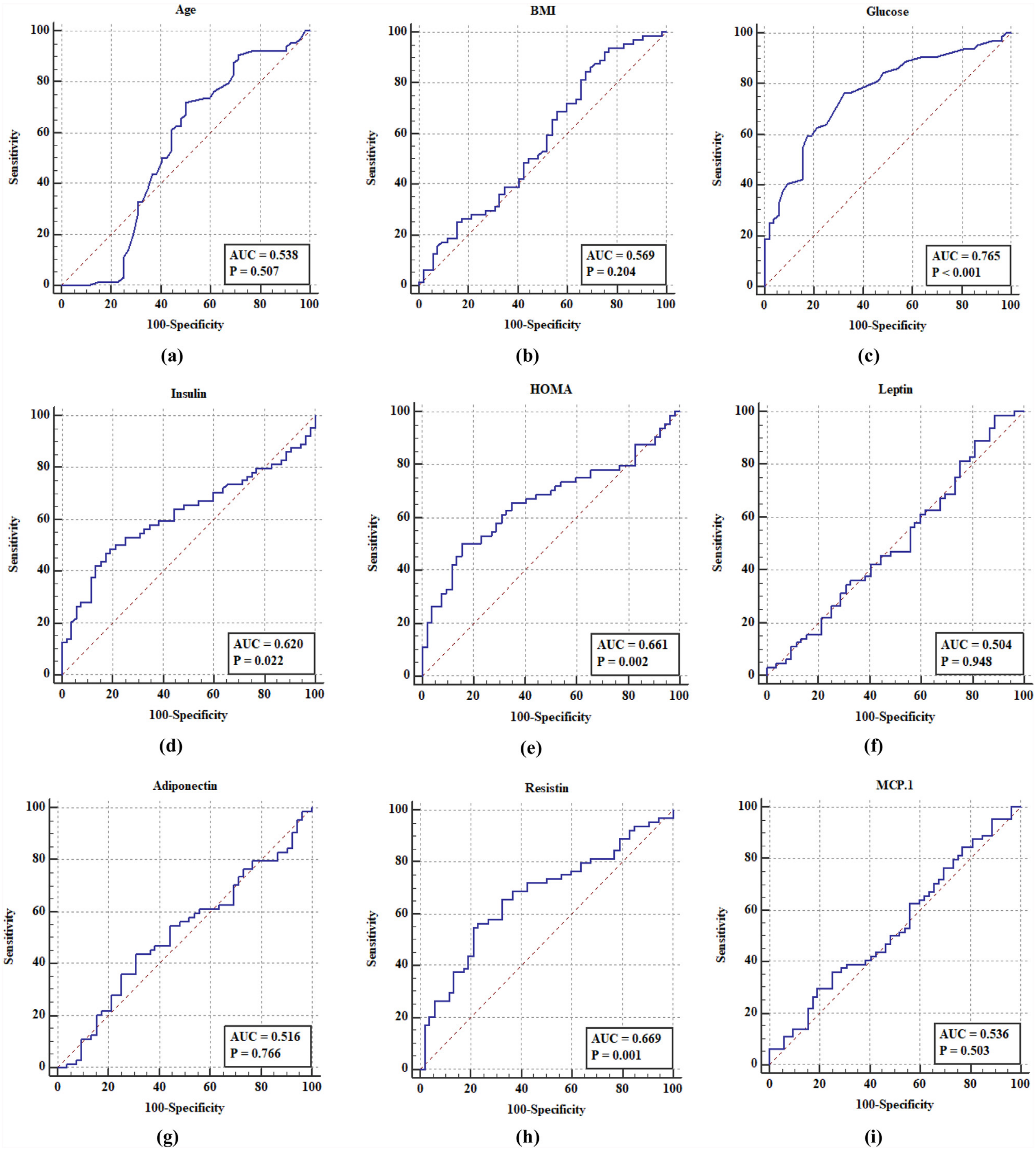
Data division protocol	Classification technique	Performance measures (average value of 5 iterations)			
		Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
5-fold	Linear SVM	71.552	70.938	72.308	0.780
	Quadratic SVM	78.621	85.000	70.769	0.820
	Cubic SVM	77.931	86.875	66.923	0.820
	Fine Gaussian SVM	75.172	88.438	58.846	0.790
	Medium Gaussian SVM	**81.897**	**90.938**	**70.769**	**0.840**
	Course Gaussian SVM	69.483	87.188	47.692	0.760
10-fold	Linear SVM	70.637	69.891	71.538	0.790
	Quadratic SVM	78.588	84.648	71.154	0.820
	Cubic SVM	78.586	84.955	70.769	0.830
	Fine Gaussian SVM	76.339	88.398	61.538	0.800
	Medium Gaussian SVM	**82.388**	**90.913**	**71.923**	**0.840**
	Course Gaussian SVM	66.608	72.540	55.769	0.770
Hold-out	Linear SVM	72.105	74.286	69.412	0.780
	Quadratic SVM	77.368	83.810	69.412	0.780
	Cubic SVM	73.158	78.095	67.059	0.760
	Fine Gaussian SVM	77.895	92.381	60.000	0.810
	Medium Gaussian SVM	**83.684**	**92.381**	**72.941**	**0.810**
	Course Gaussian SVM	70.526	84.762	52.941	0.750

3.6. *A note on sensitivity and specificity*

Sensitivity and specificity are important measures used for evaluation of different classifier models. Sensitivity can be defined as percentage of malignant samples correctly classified by the machine learning model while specificity is defined as percentage of correctly classified benign samples. Analyzing the results of Tables 6–11, it is found that the best combination of sensitivity and specificity is achieved by Medium *K*-NN classifier under hold out data division protocol when glucose, age and resistin were supplied as input to the classifier model. The values of sensitivity and specificity obtained were 95.238% and 88.235%, respectively. It is also observed that, for most of the feature combinations, sensitivity is high while specificity is low i.e. the classifier model is more accurate in categorizing malignant samples while less accurate in detecting benign samples.

3.7. *Results of ROC analysis*

The TPR and TNR are important measures to assess the diagnostic capability of features and classifiers. Receiver



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 5 – ROC plots for different features.

operating characteristics (ROC) is a plot between TPR and TNR and area under ROC (AUC) serves as a combined measure of sensitivity and specificity. In this study, ROC is plotted for each variable and AUC is calculated using MedCalc software. Fig. 5 shows the corresponding results of ROC analysis. It is found that highest AUC of 0.765 is achieved by glucose while lowest AUC of 0.504 is achieved by leptin. However, combination of these features may results in improved classifier performance and higher AUC as verified in results of Tables 6–11. The highest AUC value of 0.917 is obtained when age, glucose and resisting are used for breast cancer classification using Medium *K*-NN classifier under hold out data division protocol.

Table 10 – Performance of various SVM based classifiers using best subset selected by correlation based wrapper feature selection (CFS) and top 3 features selected by Robust rank aggregation (RRA) under different data division protocols.

Data division protocol	Classification technique	Performance measures (average value of 5 iterations)			
		Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
5-fold	Linear SVM	71.552	74.375	68.077	0.750
	Quadratic SVM	70.172	78.125	60.385	0.750
	Cubic SVM	62.241	68.125	55.000	0.670
	Fine Gaussian SVM	69.483	80.938	55.385	0.680
	Medium Gaussian SVM	**74.655**	**83.438**	**63.846**	**0.790**
	Course Gaussian SVM	64.483	83.438	41.154	0.700
10-fold	Linear SVM	71.034	73.438	68.077	0.760
	Quadratic SVM	69.719	76.473	61.403	0.750
	Cubic SVM	60.172	64.688	54.615	0.670
	Fine Gaussian SVM	70.862	81.875	57.308	0.690
	Medium Gaussian SVM	**73.621**	**82.188**	**63.077**	**0.790**
	Course Gaussian SVM	68.675	83.438	49.927	0.730
Hold-out	Linear SVM	64.737	70.476	57.647	0.760
	Quadratic SVM	71.579	78.095	63.529	0.770
	Cubic SVM	66.842	72.381	60.000	0.690
	Fine Gaussian SVM	70.526	81.905	56.471	0.710
	Medium Gaussian SVM	**75.263**	**80.000**	**69.412**	**0.800**
	Course Gaussian SVM	60.526	87.619	27.059	0.680

Table 11 – Impact of age, glucose and resistin on breast cancer prediction using other classifiers.

Data division protocol	Classification technique	Performance measures			
		Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
5-fold	Naïve Bayesian	65.471	52.821	80.545	0.667
	Linear discriminant	70.690	67.188	75.000	0.711
	Quadratic discriminant	65.517	51.563	82.692	0.671
	Logistic regression	74.138	75.000	73.077	0.740
	Fine K-NN	75.862	82.813	67.308	0.751
	Medium K-NN	79.310	82.813	75.000	0.789
	Coarse K-NN	55.172	100.000	0.000	0.500
	Cosine K-NN	67.241	56.250	80.769	0.685
	Cubic K-NN	78.448	81.250	75.000	0.781
	Weighted K-NN	**81.034**	89.063	71.154	0.801
	Random forest	77.586	81.250	73.077	0.772
10-fold	Naïve Bayesian	62.955	47.857	80.667	0.643
	Linear discriminant	72.414	67.188	78.846	0.730
	Quadratic discriminant	66.379	51.563	84.615	0.681
	Logistic regression	72.414	75.000	69.231	0.721
	Fine K-NN	78.448	81.250	75.000	0.781
	Medium K-NN	81.897	89.063	73.077	0.811
	Coarse K-NN	55.172	100.000	0.000	0.500
	Cosine K-NN	68.966	59.375	80.769	0.701
	Cubic K-NN	**83.621**	90.625	75.000	0.828
	Weighted K-NN	82.759	89.063	75.000	0.820
	Random forest	78.448	82.813	73.077	0.779
Hold-out	Naïve Bayesian	73.684	57.143	94.118	0.756
	Linear discriminant	73.684	76.190	70.588	0.734
	Quadratic discriminant	68.421	47.619	94.118	0.709
	Logistic regression	68.421	66.667	70.588	0.686
	Fine K-NN	84.211	100.000	64.706	0.824
	Medium K-NN	**92.105**	**95.238**	**88.235**	**0.917**
	Coarse K-NN	55.263	100.000	0.000	0.500
	Cosine K-NN	81.579	71.429	94.118	0.828
	Cubic K-NN	89.474	95.238	82.353	0.888
	Weighted K-NN	86.842	95.238	76.471	0.859
	Random forest	76.316	75.000	78.571	0.768

3.8. Summary of results, limitations and future scopes of the proposed study

The development in medical information systems is playing a significant part in medicine and biology. The core applications of artificial intelligence and machine learning techniques in medicine include objective analysis of medical data, computer aided diagnosis, treatment planning, reducing observational errors, assisting clinicians for faster interpretations using objective evidences, reducing inter- and intra-observational variations, providing medical facilities in low resource settings through telemedicine etc. In this study a recently available database of clinical and anthropometric measurements was utilized for breast cancer risk prediction [20]. Nine features namely glucose, HOMA, insulin, resistin, BMI, MCP.1, Age, adiponectin and leptin were used. Compared to previous study by Patrício et al. [20], this study evaluated the measurements using cross validation data division protocols and features were evaluated using eight different feature selection techniques and various popular classifiers.

Initially, all the nine features were used for classification. However, as a result of statistical significance analysis in Table 4, it was found that only four features namely glucose, insulin, HOMA and resistin were found to be statistically significant with p-value <0.05. On contradiction, results of feature selection techniques in Table 5 revealed that age can also be a significant feature for breast cancer classification. Thus, experiments were conducted initially for all possible combinations including top 5, top 4, top 3, top 2 etc. It was found that top three features i.e. age, glucose and resistin selected by 1R and Relief-F (RF) achieved highest classification accuracy among all the possible combinations. Thus, for fair comparison between different feature selection techniques, top three features selected by them were used for classification. It was also observed that when only top 2 features were considered, there is drop in classification accuracy. Hence, top 3 features were selected for each feature selection algorithms. However, since number of samples used in this work was limited, more study is required to confirm these findings on larger and multi-centric database.

Among various feature selection techniques, features selected by 1R and RF outperform others. Top three features selected by these techniques were glucose, age and Resistin. Results indicate that when these three features were used for classification, the accuracy of the classifier reaches 92.105% under hold out data division protocol which is even higher that that using all 9 features. This further indicates that insignificant and irrelevant features may misguide the classifier model thereby deteriorating its overall performance.

Among different classifiers, medium K-NN classifier outperforms others under different data division protocols followed by medium Gaussian SVM. In future, other classifiers based on extreme learning, ensemble methods, majority voting etc. can be implemented and evaluated. The results also indicate that the specificity of the proposed model is low. The clinical implication of this would be increase in number of unnecessary biopsies. Improved database with more number of pathological features can be developed and evaluated in future to improve the different performance measures including specificity so that number of unnecessary biopsies can be avoided.

4. Conclusions

This study aimed at evaluating the capability of clinical and anthropometric measurements for detecting presence of breast cancer using a database of 116 instances. Various components of machine learning paradigm like feature selection, cross validation, classification and performance evaluation were implemented and examined. Results of statistical significance analysis and feature selection techniques indicate that measures like age, glucose, insulin, HOMA and resistin have potential to be used as reliable biomarkers for detection of breast cancer. However, more study is required to confirm these findings on larger and multi-centric database of clinical and anthropometric measurements with more number of variables. Large open source databases of clinical, molecular and imaging biomarkers is needed in future to evaluate the performance of machine learning techniques in guiding breast cancer research. In future, with a larger database, the performance of techniques used in this study can be compared with the performance of the advanced classification techniques like deep neural network.

Author's contribution

The paper is singly authored by Dr. Bikesh Kumar Singh. The whole work is implemented and drafted by Dr. Bikesh Kumar Singh.

REFERENCES

[1] American Cancer Society. Cancer Facts & Figures 2016; 2016.

[2] Cancer Prevention and Early Detection: Facts and Figures 2012. American Cancer Society Cancer Action Network; 2012.

[3] Singh BK, Verma K, Thoke AS. A dual feature selection approach for classification of breast tumors in ultrasound images using ANN and SVM. Artif Intell Syst Mach Learn 2015;7(3):78–84.

[4] Singh BK, Verma K, Thoke AS. Fuzzy cluster based neural network classifier for classifying breast tumors in ultrasound images. Expert Syst Appl 2016;66:114–23.

[5] Singh BK, Verma K, Panigrahi L, Thoke AS. Integrating radiologist feedback with computer aided diagnostic systems for breast cancer risk prediction in ultrasonic images: an experimental investigation in machine learning paradigm. Expert Syst Appl 2017;90:209–23.

[6] Singh BK, Verma K, Thoke AS, Suri JS. Risk stratification of 2D ultrasound-based breast lesions using hybrid feature selection in machine learning paradigm. Measurement 2017;105:146–57.

[7] Panigrahi L, Verma K, Singh BK. Ultrasound image segmentation using a novel multi-scale Gaussian kernel fuzzy clustering and multi-scale vector field convolution. Expert Syst Appl 2019;115:486–98.

[8] Opstal-van Winden AW, Rodenburg W, Pennings JL, van Oostrom C, Beijnen JH, Peeters PH, van Gils CH, de Vries A. A bead-based multiplexed immunoassay to evaluate breast cancer biomarkers for early detection in pre-diagnostic serum. Int J Mol Sci 2012;13(10):13587–604.

[9] Santillán-Benítez JG, Mendieta-Zerón H, Gómez-Oliván LM, Torres-Juárez JJ, González-Bañales JM, Hernández-Peña LV,

Ordóñez-Quiroz A. The tetrad BMI, leptin, leptin/adiponectin (L/A) ratio and CA 15-3 are reliable biomarkers of breast cancer. J Clin Lab Anal 2013;27(1):12–20.

[10] Dalamaga M, Sotiropoulos G, Karmaniolas K, Pelekanos N, Papadavid E, Lekka A. Serum resistin: a biomarker of breast cancer in postmenopausal women? Association with clinicopathological characteristics, tumor markers, inflammatory and metabolic parameters. Clin Biochem 2013;46(7–8):584–90.

[11] Kloten V, Becker B, Winner K, Schrauder MG, Fasching PA, Anzeneder T, Veeck J, Hartmann A, Knüchel R, Dahl E. Promoter hypermethylation of the tumor-suppressor genes ITIH5, DKK3, and RASSF1A as novel biomarkers for blood-based breast cancer screening. Breast Cancer Res 2013;15(1):R4.

[12] Zhu Q, Wang L, Tannenbaum S, Ricci A, DeFusco P, Hegde P. Pathologic response prediction to neoadjuvant chemotherapy utilizing pretreatment near-infrared imaging parameters and tumor pathologic criteria. Breast Cancer Res 2014;16(5):456.

[13] Zheng B, Yoon SW, Lam SS. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Syst Appl 2014;41(4):1476–82.

[14] Provatopoulou X, Georgiou GP, Kalogera E, Kalles V, Matiatou MA, Papapanagiotou I, Sagkriotis A, Zografos GC, Gounaris A. Serum irisin levels are lower in patients with breast cancer: association with disease diagnosis and tumor characteristics. BMC Cancer 2015;15(1):898.

[15] Assiri AM, Kamel HF. Evaluation of diagnostic and predictive value of serum adipokines: leptin, resistin and visfatin in postmenopausal breast cancer. Obes Res Clin Pract 2016;10(4):442–53.

[16] Lee E, Moon A. Identification of biomarkers for breast cancer using databases. J Cancer Prev 2016;21(4):235.

[17] Alickovic´ E, Subasi A. Breast cancer diagnosis using GA feature selection and rotation forest. Neural Comput Appl 2017;28(4):753–63.

[18] Choi J, Park S, Yoon Y, Ahn J. Improved prediction of breast cancer outcome by identifying heterogeneous biomarkers. Bioinformatics 2017;33(22):3619–26.

[19] Nicolini A, Ferrari P, Duffy MJ. Prognostic and predictive biomarkers in breast cancer: past, present and future. Semin Cancer Biol 2018;52:56–73.

[20] Patrício M, Pereira J, Crisóstomo J, Matafome P, Gomes M, Seiça R, Caramelo F. Using resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer 2018;18(1):29.

[21] Phillips M, Cataneo RN, Cruz-Ramos JA, Huston J, Ornelas O, Pappas N, Pathak S. Prediction of breast cancer risk with volatile biomarkers in breath. Breast Cancer Res Treat 2018;1–8.

[22] Chen YZ, Kim Y, Soliman H, Ying G, Lee JK. Single drug biomarker prediction for ER− breast cancer outcome from chemotherapy. Endocr Relat Cancer 2018;25(6):595–605.

[23] Besutti G, Iotti V, Rossi PG. Molecular imaging biomarkers for breast cancer risk and personalized screening. Transl Cancer Res 2018;7(5):1319–25.

[24] Kyrochristos ID, Ziogas DE, Lykoudis EG, Roukos DH. Breast cancer genome analysis in time and space: biomarker development strategy. Biomark Med 2018;12(6):547–50.

[25] Feng W, Li Y, Chu J, Li J, Zhang Y, Ding X, Fu Z, Li W, Huang X, Yin Y. Identification of tRNA-derived small noncoding RNA s as potential biomarkers for prediction of recurrence in triple-negative breast cancer. Cancer Med 2018;7(10):5130–44.

[26] Weaver O, Leung JW. Biomarkers and imaging of breast cancer. Am J Roentgenol 2018;210(2):271–8.

[27] Landau S, Everitt BS. A handbook of statistical analyses using SPSS. Chapman & Hall/CRC; 2004.

[28] Chandrashekar G, Sahin F. A survey on feature selection methods. Comput Electr Eng 2014;40(1):16–28.

[29] Setiono R, Liu H. Improving backpropagation learning with feature selection. Appl Intell 1996;6(2):129–39.

[30] Zhao Z, Morstatter F, Sharma S, Alelyani S, Anand A, Liu H. Advancing feature selection research. ASU Feature Sel Repos 2010;1–28.

[31] Holte RC. Very simple classification rules perform well on most commonly used datasets. Mach Learn 1993; 11(1):63–90.

[32] Robnik-Šikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. Mach Learn 2003; 53(1–2):23–69.

[33] Yu L, Liu H. Feature selection for high-dimensional data: a fast correlation-based filter solution. Proceedings of the 20th International Conference on Machine Learning (ICML-03); 2003. p. 856–63.

[34] Hall MA. Correlation-based feature selection for machine learning; 1999.

[35] Kolde R, Laur S, Adler P, Vilo J. Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics 2012;28(4):573–80.

[36] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97.

[37] Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. Proceedings of the Fifth Annual Workshop on Computational Learning Theory. ACM; 1992. p. 144–52.

[38] Tharwat A. Linear vs. quadratic discriminant analysis classifier: a tutorial. Int J Appl Pattern Recognit 2016; 3(2):145–80.

[39] Li T, Zhu S, Ogihara M. Using discriminant analysis for multi-class classification: an experimental investigation. Knowl Inf Syst 2006;10(4):453–72.

[40] Kalmegh SR. Comparative analysis of weka data mining algorithm random forest, random tree and lad tree for classification of indigenous news data. Int J Emerg Technol Adv Eng 2015;5(1):507–17.

[41] Cappelletti V, Appierto V, Tiberio P, Fina E, Callari M, Daidone MG. Circulating biomarkers for prediction of treatment response. J Natl Cancer Inst Monogr 2015; 2015(51):60–3.

[42] Novakovic´ J, Strbac P, Bulatovic D. Toward optimal feature selection using ranking methods and classification algorithms. Yugosl J Oper Res 2016;21(1).

[43] Isaac ER. Test of hypothesis – concise formula summary. Available from: https://www.researchgate.net/profile/Ebenezer_Isaac/publication/283318687_Test_of_Hypothesis_-_Concise_Formula_Summary/links/5632e74c08aefa44c3685cd7/Test-of-Hypothesis-Concise-Formula-Summary.pdf [accessed 09.03.19].