Predicting breast cancer recurrence using principal component analysis as feature extraction: an unbiased comparative analysis

Zuhaira Muhammad Zain [a,1,*], Mona Alshenaifi [a,2], Abeer Aljaloud [a,3], Tamadhur Albednah [a,4], Reham Alghanim [a,5], Alanoud Alqifari [a,6], Amal Alqahtani [a,7]

[a] Information Systems Department, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia
[1] zmzain@pnu.edu.sa; [2] 434008386@pnu.edu.sa; [3] 435005150@pnu.edu.sa; [4] 435010107@pnu.edu.sa; [5] 4345001378@pnu.edu.sa; [6] 435006088@pnu.edu.sa; [7] 434002913@pnu.edu.sa
* corresponding author

ARTICLE INFO

Article history
Received January 29, 2020
Revised October 29, 2020
Accepted November 6, 2020
Available online November 30, 2020

Keywords
Breast cancer recurrence
Data mining
Feature extraction
Machine learning
Principal component analysis

ABSTRACT

Breast cancer recurrence is among the most noteworthy fears faced by women. Nevertheless, with modern innovations in data mining technology, early recurrence prediction can help relieve these fears. Although medical information is typically complicated, and simplifying searches to the most relevant input is challenging, new sophisticated data mining techniques promise accurate predictions from high-dimensional data. In this study, the performances of three established data mining algorithms: Naïve Bayes (NB), k-nearest neighbor (KNN), and fast decision tree (REPTree), adopting the feature extraction algorithm, principal component analysis (PCA), for predicting breast cancer recurrence were contrasted. The comparison was conducted between models built in the absence and presence of PCA. The results showed that KNN produced better prediction without PCA (F-measure = 72.1%), whereas the other two techniques: NB and REPTree, improved when used with PCA (F-measure = 76.1% and 72.8%, respectively). This study can benefit the healthcare industry in assisting physicians in predicting breast cancer recurrence precisely.

1. Introduction

Breast cancer is the most prevalent cancer among women, affecting 2.1 million women per year. It also causes the largest number of cancer-related deaths among women. The World Health Organization (WHO) also announced that an estimated 627,000 people died of breast cancer in 2018, around 15% of all cancer deaths among women. While breast cancer rates are higher among women in more developed countries, rates rise in almost every region of the world [1].

Breast cancer recurrence is one of the biggest challenges a patient has to face and is one of the issues that impact their living standards. Breast cancer recurrence refers to breast cancer reoccurring in a woman whose former cancer was remediated. According to Pan *et al.* [2] even 20 years after a diagnosis, women with a type of breast cancer fueled by estrogen still face a substantial risk of cancer returning or spreading. The prediction is challenging because the recurrence data is rarely recorded in most breast cancer datasets. An accurate and timely prediction is essential because it helps physicians make a decision and supports more personalized patient therapy.

Knowledge discovery in databases (KDD) methods offer a stimulating prospect to scrutinize this kind of problems driven by data. Data mining, an important KDD subset, remains an iterative procedure in the hunt for current, useful, as well as critical data in enormous data amounts, also called high-dimensional data [3]. Multiple health-related illnesses, key among them being breast cancer [4]–[11] diabetes [4][12][13], and oral cancer [12], in addition to cardiovascular diseases [13]–[16], have been effectively diagnosed and predicted by utilizing data mining, as well as machine learning procedures. These successful studies encourage the data mining application in predicting breast tumor recurrence, therefore driving the foundation of this study.

In the last few years, different types of high-dimensional information have been generated by developing high-throughput technologies, especially those associated with the manifestation of disease and the control of tumor recurrence. It is a challenge to get insights from high-dimensional data. High-dimensional data have to be transformed into low-dimensional data by operating reduction techniques. Dimensionality reduction enables high-dimensional data to be classified, visualized, communicated and stored.

The medical data dimensions contain a number of features, and every feature comprises various types of values. Data quality problems consist of missing or redundant data, outliers, noise, and biased or unrepresentative data entries [17]. To focus on data preparation, preprocessing stages should be used to increase the suitability of raw data for analysis. Additionally, medical data entries are usually complex and suffer from the challenge of high dimensionality. It is difficult to reduce the dataset used in the prediction manually, but a feature extraction technique can be used to solve this. Some popular feature extraction techniques include principal component analysis (PCA), independent component analysis, linear discriminant analysis, locally linear embedding, t-distributed stochastic neighbor embedding, and autoencoders [18]. Among them, PCA is widely used in breast cancer prediction [19]. Besides, it is the most appropriate approach that can be applied when there is a need to minimize the number of variables. However, it cannot specify which variable to keep in consideration. Also, PCA works best on datasets with three or higher dimensions of numeric variables.

Moreover, it aims to reduce feature dimension by capturing as much information as possible with high explained variance and minimizing information loss at the same time. With emerging techniques in data mining, the production of accurate predictions is promising. However, feature extraction alone is not sufficient to predict breast cancer recurrence.

Some classification algorithms such as K-nearest neighbors (KNN), Naïve Bayes (NB), and fast decision tree (REPTree) need to be applied to classify whether patients have breast cancer recurrence or not. These classifiers have been used for many healthcare data prediction [7][9][13][15][20][21]. However, the three popular classifiers have not been combined with PCA as feature extraction in predicting breast cancer recurrence. Another issue in machine learning studies is regarding the performance metrics used. Most of the studies usually used accuracy on the models' performance evaluation while there are many more performance metrics that can be used to measure the performance of machine learning classifiers like incorrectly classified instances, Cohen's kappa, recall, precision, and F-measure.

This study proposed a PCA technique to reduce the Wisconsin Prognostic Breast Cancer dataset's high dimensionality to tackle the aforementioned drawbacks. KNN, NB, and REPTree were used as classifiers in the prediction models. Performance metrics, such as incorrectly classified instances, Cohen's kappa, recall, precision, and F-measure, were applied in addition to accuracy during the comparative analysis to evaluate the distinction between the performance demonstrated by PCA models and non-PCA models.

Additional sections on the manuscript are organized accordingly. Section 2 defines the PCA feature extraction, NB, KNN, and REPTree classifiers used in this study. Then, Section 3 clarifies every phase of the research methodology. Section 4 examines the findings of this research. Finally, Section 5 sets out the conclusions and emphasizes the extent of future activities.

2. Related Work

2.1. Feature Extraction Using PCA

Feature extraction is the procedure by which irrelevant, less relevant, or redundant dimensional attributes are identified and disregarded within a given dataset [22][23] that transforms data in high-dimensional space to less-dimensional space. These methods usually are denoted as preprocess to machine learning algorithms (MLA) for pattern recognition and prediction [24]. PCA is one of the feature extraction approaches.

Using PCA makes it possible to reduce the number of variables in a multivariate dataset, preserving as much variation as possible in the dataset. Such minimization is accomplished through the employment of distinct p variables, namely, T_1, T_2 , T_3 . . .,T_p and finding the groupings of the variables to generate uncorrelated principal elements (*PCs*) PC_1, PC_2, PC_3 . . ., PC_p. The aforementioned *PCs* are also known as eigenvectors. Notably, correlation deficiencies make up an invaluable property because it indicates that different "dimensions" are computed within the data through the *PC*. However, *PCs* are arranged in such a way that PC_1 shows the greatest variation, while PC_2 shows the second greatest variation, and the subsequent *PCs* reduce their variation uniformly. Basically, $var(PC_1)$ is greater than or equal to $var(PC_2)$, $var(PC_2)$ is greater than or equal to $var(PC_3)$, and $var(PC_3)$ is greater than or equal to $var(PC_p)$. In this scenario, $var(PC_i)$ represents the PC_i variation within the relevant dataset. Meanwhile, $var(PC_i)$ may be denoted as PC_i's eigenvalue.

The PCA algorithm starts with calculating the mean for each feature. The mean value is then subtracted from the original data to the new centralized data, and it decomposes the covariance matrix of the data. Afterward, the covariance matrix of data points is calculated, and its eigenvectors and corresponding eigenvalues are solved. Next, the eigenvectors, according to their eigenvalues, are sorted in decreasing order. Choosing the first k (number of components) eigenvectors will yield the new k dimensions. Finally, PCA would transform the original dimensional data points into the new reduced dimensions.

Several studies have utilized PCA as the feature extraction method on healthcare data, especially the Wisconsin Breast Cancer dataset. For example, in [8], PCA was combined with a differential evolution support vector machine to improve the cancer detection ability with 97.64% accuracy. Hasan and Tahir [25] applied PCA as feature extraction and the artificial neural network as a classifier to enhance benign or malignant classification. Their method was found to discriminate between normal and breast cancer patients with 95.68% testing accuracy. Jamal *et al.* [19] implemented PCA with Support Vector Machine and Extreme Gradient Boosting in predicting breast cancer. Jhajharia *et al.* [26] conducted a study where PCA is applied together with artificial neural networks with 98.39% accuracy. Uzer *et al.* [27] conducted another successful study on breast cancer prediction. First, they selected important features by using sequential forward selection (SFSP) and sequential backward selection (SBSP) algorithms. The selected features from both algorithms were then fed to the PCA to reduce the dimensionality.

The new feature set was then used as an input for the neural network classifier. Their study achieved 98.57% and 97.57% accuracy for SFSP and SBSP, respectively. A recent study conducted by Bian et al. [28] proposed a new breast cancer prediction approach. They employed random forest as a feature selection to select a set of important features. These features were passed to PCA to reduce data dimensionality. The new feature set of seven principal components was finally fed to the extreme learning machine classification model with different activation functions. Their proposed model achieved 98.75% accuracy. Another study conducted by Roopa and Asha [29] achieved 96.07% accuracy using PCA with wrapper and linear regression algorithms in tuberculosis diagnosis. All of these studies show that applying PCA reduces the dimension of the dataset and increases the performance of the classifiers. However, none of these studies tested PCA with the three famous classifiers, namely, NB, REPTree, and KNN, in breast cancer recurrence detection.

2.2. Classification Algorithms

This subsection describes the three well-known classification algorithms used in healthcare: NB, REPTree, and KNN.

2.3. NB

NB denotes a Bayes' theorem classification method that contains a supposition of autonomy between the classification algorithms. This classifier supposes that there is no connection between a particular feature and the presence of any other feature within a class. It is worth noting that the model is invariably easy to develop and is very capable of enormous datasets. NB is known for its simplicity and outstanding classification processes [30]. It also performs very well in multiclass predictions, as an easy and fast predictor of single class test sets. If the independence assumption is retained, an NB classifier performs better than logistic regression and takes less training data. Besides, it performs well in categorical input compared with the normalized bell curve. Several studies have been carried out using the NB algorithm on healthcare data [7][13][15], and findings confirm that it is a good classifier in predicting healthcare-related cases.

2.4. REPTree

The REPTree classifier denotes a fast decision tree learning system constructed from the concept of calculating the data gain with entropy and reducing errors resulting from variance [31]. It was proposed in 2011 [32]. It uses the logic of a regression tree and produces several trees in modified iterations. The best tree of the spawned trees will then be selected. This algorithm uses variance and information gain to build the regression/decision tree. Further, this algorithm uses a back-fitting method to prune the tree with reduced error pruning. It sorts numerical attribute values once at the start of the preparation of the model. This algorithm also addresses missing values, as in the C4.5 algorithm, by dividing the corresponding scenarios into pieces [33]. This study [21] reported that REPTree performs well in classifying healthcare data.

2.5. KNN

The KNNs denote a supervised classifier that selects the k nearest neighbor associated with a particular point by minimizing a similarity measure, the Mahalanobis distance or Euclidean distance [20]. KNN calculates its closeness to the outstanding (labeled) instances and establishes its k-nearest neighbor and their respective labels to determine the class of an unlabeled example. The unlabeled object is subsequently categorized either by a majority vote by the neighborhood's dominant category or through a predominantly weighted majority whereby points nearer to the unlabeled object are given greater weight. KNN is considered a good classifier in healthcare data recognition and prediction [9][20].

Nevertheless, to the best of our knowledge, the aforementioned three prominent classifiers have not been combined with PCA as feature extraction in predicting breast cancer recurrence.

3. Method

As shown in Fig. 1, the research method is broken down into five phases: data acquisition, preprocessing of data, model construction without PCA, model construction with PCA, and model comparison.

3.1 Phase One: Data Acquisition

In this phase, the study's pertinent data is acquired from UCI's public repository [34]. The Wisconsin Prognostic Breast Cancer (WPBC) dataset consists of 34 predictors/independent features and output/dependent features, in addition to 198 records. Individual record denotes further examination data for a breast cancer case for Dr. Wolberg's patients since 1984. There are 151 nonrecurrence cases and 47 recurrence cases. There exist missing values within the lymph node status feature in four cases.

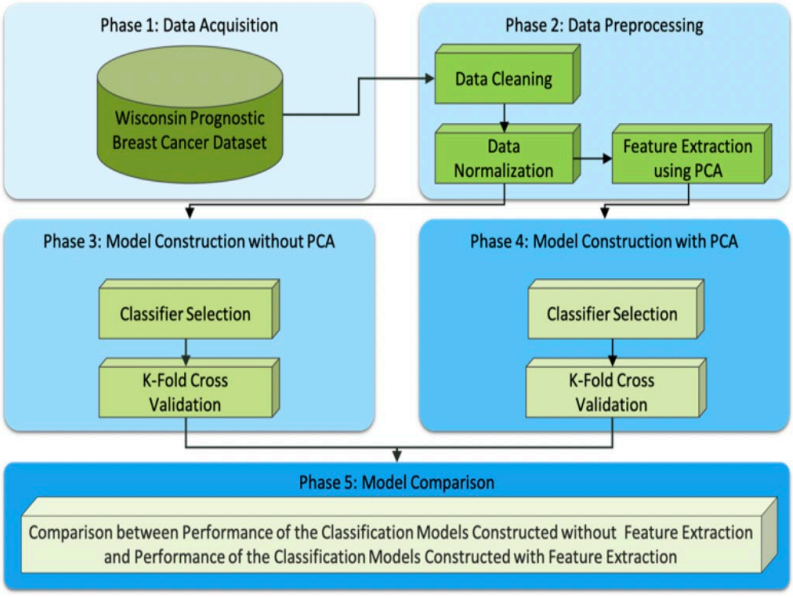


Fig. 1. Method of Research

3.2 Phase Two: The Preprocessing of Data

The number of processes in the data preprocessing phase depends on developing a model with or without feature extraction. To develop a model without feature extraction, the data preprocessing phase comprises two processes: data cleaning and normalization. The second model adds a third step of feature extraction.

Over the course of the process of cleaning data, the missing datasets within the lymph node status were given the most probable value by employing the *ReplaceMissingValues* filter within the Weka. Notably, the ID feature was deleted from the dataset because it will not affect the outcome. This process reduced the number of predictor features to 33. Data normalization is useful when the dataset has varying scales. It is worth noting that the data acquired in Phase 1 was rescaled to a range between 0 and 1 by utilizing the *Normalize* filter within Weka to attain similar value ranges for every feature.

In the feature extraction process, PCA was applied in order to reduce the feature dimension. Reasons for opting for PCA are as follows: (1) PCA aims to capture as much information as possible with high explained variance, unlike any other algorithms that only select several important features that cause information loss; (2) PCA works best on datasets with three or higher dimensions, such as the WPBC dataset, which consists of 33 attributes, and since it has the highest dimensions, it is increasingly difficult to interpret the result; and (3) PCA is ideal for use on a dataset of numeric variables such as WPBC. When applying PCA, it is best to choose a few principal components with variance covered as high as possible. In Weka, we just need to set the variance covered to 0.95. The PCA algorithm automatically selected an optimal number of principal components, with 13 principal components representing 33 features by minimizing information loss. In other words, by using PCA, the number of predictors has been reduced from 33 to 13 without compromising on explained variance. The spree plot in Fig. 2 shows the number of principal components selected with the proportion of variance. The red line indicates the variance covered per component, and the green line indicates the cumulative variance covered by components.

PCA also provides the principal component loading (Fig. 3). It can be inferred that the first principal component, PC1, corresponds to a measure of *0.28813Mean_Concave_points* + *0.277034Mean_Concavity* + ... + *0.0147931Lymph_node_status*. Similarly, it can be said that the second component, PC2, corresponds to a measure of *0.301744Mean_Fractal_dimension* + *0.288685Worst_Fractal_dimension* + ... – *0.00457267Worst_Texture*. PCA then computes eigenvectors that are the principal component and respective eigenvalues that apprehend the magnitude of variance. Finally, the eigenpairs were arranged to decrease the order of respective eigenvalues, and the value with the maximum value was picked. This

is the first principal component that protects the maximum information from the original data. The new data frame was created from 13 principal components and their eigenvalues. Table 1 presents the sample of the first ten rows of the data frame that will be used as an input in Phase 4.

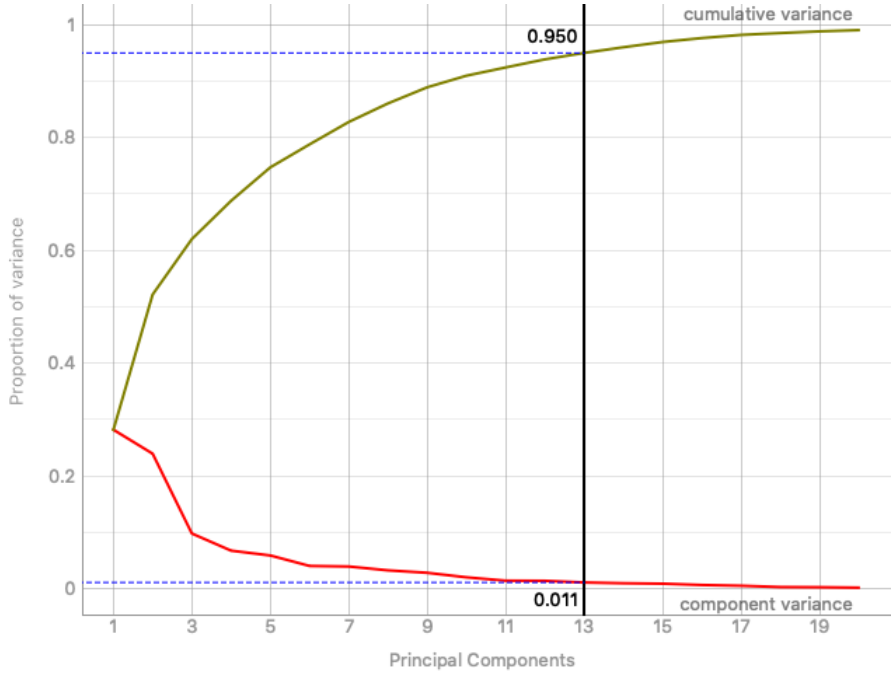


Fig. 2. Spree plot of the proportion of variance for each principal component

PC1	Weight
Mean_Concave_points	0.28813
Mean_Concavity	0.277034
Worst_Perimeter	0.247672
Mean_Perimeter	0.243186
Area_SE	0.240875
Perimeter_SE	0.239893
Mean_Area	0.231804
Radius_SE	0.230617
Worst_Radius	0.230128
Worst_Concave_points	0.229777
Mean_Radius	0.228437
Mean_Compactness	0.227836
Worst_Area	0.224842
Concavity_SE	0.194206
Concave_points_SE	0.174857
Compactness_SE	0.172751
Worst_Concavity	0.160329
Fractal_dimension_SE	0.147904
Mean_Symmetry	0.131052
Mean_Smoothness	0.123825
Worst_Compactness	0.111315
Symmetry_SE	0.110581
Smoothness_SE	0.0832344
Time	-0.0763553
Texture_SE	0.07454
Mean_Fractal_dimension	0.0714788
Worst_Fractal_dimension	0.035333
Worst_Symmetry	0.0324785
Worst_Smoothness	0.0269383
Mean_Texture	0.0228055
Tumor_size	0.0211516
Worst_Texture	-0.0168169
Lymph_node_status	0.0147931

PC2	Weight
Mean_Fractal_dimension	0.301744
Worst_Fractal_dimension	0.288685
Worst_Compactness	0.251012
Worst_Smoothness	0.238637
Mean_Smoothness	0.226488
Mean_Radius	-0.212552
Fractal_dimension_SE	0.212479
Worst_Concavity	0.212272
Mean_Area	-0.212082
Worst_Symmetry	0.21187
Mean_Compactness	0.208479
Mean_Symmetry	0.205393
Worst_Radius	-0.199426
Worst_Area	-0.198406
Compactness_SE	0.196464
Mean_Perimeter	-0.19373
Worst_Perimeter	-0.176613
Area_SE	-0.153785
Concavity_SE	0.145219
Radius_SE	-0.126365
Worst_Concave_points	0.122082
Time	0.121391
Symmetry_SE	0.115494
Perimeter_SE	-0.109864
Mean_Concavity	0.105022
Smoothness_SE	0.0899702
Tumor_size	-0.0707345
Mean_Texture	-0.0589697
Concave_points_SE	0.0439191
Mean_Concave_points	0.0156933
Lymph_node_status	-0.0148034
Texture_SE	-0.00765185
Worst_Texture	-0.00457267

...

PC13	Weight
Concave_points_SE	-0.514549
Worst_Concave_points	-0.34499
Mean_Texture	-0.31364
Time	-0.311821
Texture_SE	0.2715
Tumor_size	-0.225873
Fractal_dimension_SE	0.198634
Lymph_node_status	0.176572
Worst_Area	0.154187
Worst_Fractal_dimension	0.148611
Smoothness_SE	0.122769
Mean_Concave_points	-0.122358
Area_SE	0.122228
Mean_Symmetry	-0.120899
Compactness_SE	0.112739
Worst_Smoothness	0.111909
Worst_Symmetry	0.102707
Mean_Area	0.10246
Worst_Radius	0.0904321
Perimeter_SE	-0.0854416
Worst_Perimeter	0.0641613
Mean_Radius	0.0625363
Mean_Smoothness	0.0622112
Worst_Texture	-0.057912
Symmetry_SE	0.0578785
Mean_Compactness	-0.0571746
Mean_Perimeter	0.0556353
Worst_Compactness	0.0531699
Mean_Fractal_dimension	0.049333
Mean_Concavity	-0.0314989
Worst_Concavity	-0.0162844
Radius_SE	-0.0117511
Concavity_SE	-0.00883397

Fig. 3. Sample of principal component loading

Table 1. Sample of 10 rows of the new data frame created from 13 principal components

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	Outcome
11.58	2.64	5.36	0.54	0.47	-1.26	-0.65	-0.26	2.02	1.74	-0.57	-1.85	2.93	0
8.49	-2.40	-3.83	0.80	1.45	-0.01	-0.33	0.91	0.73	0.17	-0.36	-1.11	0.40	1
7.92	-3.36	2.72	-1.67	-2.29	-0.44	-1.89	-2.67	-0.93	-0.58	-1.32	-0.12	-0.13	0
7.06	-2.05	0.27	-0.76	-2.53	-0.32	-1.42	-0.93	-1.45	-1.60	0.41	0.63	0.17	0
6.97	-4.44	-3.80	-0.87	0.37	-0.66	-0.55	0.29	1.36	0.67	1.00	-0.93	0.35	1
6.36	0.41	-0.20	-2.00	-0.48	-1.67	3.56	-0.14	2.45	0.07	-0.78	-0.57	-1.04	0
-6.29	-1.71	0.01	0.58	-0.56	1.16	0.05	0.37	-0.09	-1.01	0.13	0.06	0.12	0
6.15	2.37	1.46	-1.52	1.33	5.08	0.76	1.39	-1.51	0.02	-0.40	0.22	-0.62	1
-6.04	-0.55	1.40	-0.59	-0.29	0.14	-0.96	0.45	0.50	-0.68	-0.51	0.02	1.06	0
6.03	-8.58	-1.15	1.71	1.81	0.05	-1.57	-0.57	-2.73	-0.10	-1.09	1.24	1.12	1

3.3 Phase Three: Model Construction Without PCA

This stage entailed constructing classification models by employing three common classifiers, namely, KNN, NB, and REPTree, with a tenfold cross-validation test alternative by using Weka. The dataset was divided into ten pieces (folds), and each piece was then kept in turn for testing, and the remaining nine pieces were trained together. The average for ten evaluation results was calculated. After that, the classifier was invoked for the last (11th) time by Weka on the entire dataset to print out the final evaluation result.

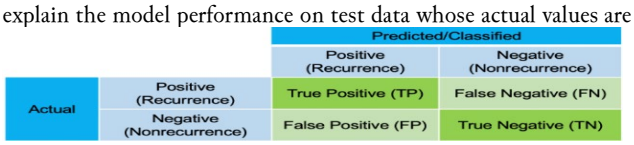
3.4 Phase Four: Model Construction With PCA

This phase entailed carrying out feature extraction by utilizing the PCA obtained from Phase 2 to minimize the dimensionality of the dataset. After that, Phase 3 was repeated to build three classification models with the reduced feature set.

3.5 Phase Five: Model Comparison

The performance of the prediction models built with and without PCA was compared in this phase. Most previous research only used one or two performance criteria, leading to bias in the result discussion. Table 2 lists the performance criteria with their descriptions.

Table 2. Performance criteria with their description

Performance Criteria	Description
Confusion matrix	A table to explain the model performance on test data whose actual values are known  • True positives refer to data points that the model determines as positive, which are actually positive (correct) • True negatives denote data points that the model determines as negative, which are actually negative (correct) • False positives refer to data points that the model determines as positive, which are really negative (incorrect) • False negatives denote data points that the model determines as negative, which are really positive (incorrect)
Accurately classified instances	• The percentage of instances accurately classified • The sum of true positive and true negative instances (TP+TN)
Inaccurately classified instances	• The percentage of instances inaccurately classified • The sum of false positive and false negative instances (FP+FN)
Kappa statistic	Denotes the estimate of how well the model can separate the instances into the right class. Notably, Cohen's kappa ≤ 1. Values < 0 serve to illustrate the ineffectiveness of the classifier. Interpretation scheme < 0 illustrates no agreement, 0–0.20 indicates slight agreement, 0.21–0.40 signifies fair agreement, 0.41–0.60 shows moderate agreement, 0.61–0.80 illustrates substantial agreement, and 0.81–1 shows close to a perfect agreement [35] Cohen's kappa = (totalAccuracy–randomAccuracy)/(1–randomAccuracy) totalAccuracy = (TP + TN)/(TP + TN + FP + FN) randomAccuracy = ((TN + TP)*(TN + FN)+(FN + TP)*(FP + TP))/(Total * Total)

Performance Criteria	Description
Precision (P)	The ability to determine only the relevant data points by a classification model Denotes the proportion of correctly forecasted positive observations in relation to the cumulative forecasted positive observations Precision = True Positives/(True Positives + False Positives)
Recall (R)	The ability to locate all the pertinent cases in a dataset by a classification model Refers to the proportion of accurately forecasted positive observations in relation to every observation within the actual category Recall = True Positive/(True Positive + False Negative)
F-Measure (F)	Refers to the harmonic average associated with precision as well as recall F = 2*(R*P)/(R+P)

4. Results and Discussion

This section examines the models constructed in Phases 3 and 4 and the results obtained in the comparative analysis of Phase 5.

4.1. Evaluation of the Models Constructed Without PCA

Table 3 lists the summary statistics for the three classification models without feature extraction. The summary shows that REPTree outperforms the other two classifiers by correctly classifying 149 (75.25%) instances. However, the negative Cohen's kappa value (–0.0198) indicates that REPTree is not an effective classifier for predicting whether a patient has breast cancer recurrence. There is slight agreement to say that NB is an effective breast cancer recurrence classifier with Cohen's kappa value of 0.1794 and fair agreement to say that KNN is an effective breast cancer recurrence classifier with Cohen's kappa value of 0.2271.

Table 3. Summary statistic for classification models without feature extraction

Classifier	Accurately Classified Instances	Inaccurately Classified Instances	Kappa Statistic
Naives Bayes	134 (67.68%)	64 (32.32%)	0.1794
REPTree	149 (75.25%)	49 (24.75%)	–0.0198
KNN	143 (72.22%)	55 (27.78%)	0.2271

The prognostic breast cancer prediction duty is an imbalanced classification issue where two classes require to be predicted, namely, recurrence and nonrecurrence, with nonrecurrence indicating the tremendous majority of the data points. The confusion matrix is displayed in Table 4 breaks down the data in Table 3 and represents the actual and predicted labels from the classification results of the three models. The TPs are correctly identified recurrence cases, and TNs are correctly identified nonrecurrence cases. Conversely, FPs are patients who would be falsely identified as recurrence cases, and FNs are patients who would be falsely identified as nonrecurrence cases. Table 4 corroborates that the REPTree model is an ineffective classifier because it can correctly identify nonrecurrence cases (TN = 149) but not the recurrence cases (TP = 0).

Table 4. Confusion matrix for classification models without feature extraction

Classifier	*Ra*	*NRb*	*Classified As*
Naïve Bayes	21 (TP)	26 (FN)	R
	38 (FP)	113 (TN)	NR
REPTree	0 (TP)	47 (FN)	R
	2 (FP)	149 (TN)	NR
KNN	19 (TP)	28 (FN)	R
	27 (FP)	124 (TN)	NR

[a.] Recurrence, [b.] Nonrecurrence

The detailed accuracy by class was then examined to inspect the performance of the three classifiers further. Table 5 represents the accuracy by class in detail for the three models without feature extraction. In this analysis, creating a balanced classification model with the optimal balance of recall and precision remains the top priority. The weighted average for recall, precision, and F-measure for two classes was calculated using (1).

$$weighted\ average = \frac{(X_{c1}*|c1|)+(X_{c2}*|c2|)}{|c1|+|c2|}, \quad (1)$$

where X can be the value for precision (P), recall (R), or F-measure (F), c1 is the number of instances in class 1, and c2 is the number of instances in class 2. Below is the example of the calculation of the weighted average for F-measure for NB:

$$weighted\ average_{\mathrm{NB}} = \frac{(F_{c1}*|c1|)+(F_{c2}*|c2|)}{|c1|+|c2|} \quad (2)$$

$$weighted\ average_{\mathrm{NB}} = \frac{(0.396*|47|)+(0.779*|151|)}{|47|+|151|}$$

$$weighted\ average_{\mathrm{NB}} = 0.688$$

F-measure is employed to determine which model is the best to classify patients into the recurrence and nonrecurrence categories. The F-measure values (0.721) signify that KNN outperforms the other two classifiers in predicting whether the patients have breast cancer recurrence.

Table 5. Detailed accuracy for classification models without feature extraction

Classifier	P	R	F	Class
Naïve Bayes	0.356	0.447	0.396	R[a]
	0.813	0.748	0.779	NR[b]
Weighted Average	0.704	0.677	0.688	
REPTree	0.000	0.000	0.000	R
	0.760	0.987	0.859	NR
Weighted Average	0.580	0.753	0.655	
KNN	0.413	0.404	0.409	R
	0.816	0.821	0.818	NR
Weighted Average	0.720	0.722	0.721	

[a.] Recurrence, [b.] Non-recurrence

4.2. Evaluation of the Models Constructed With PCA

Upon completing the feature extraction process, PCA transformed 33 correlated features into a novel set of 13 linearly uncorrelated principal components that captured over 95% of the training dataset's initial variance.

The results are shown in Table 6 imply that NB outperforms the other two classification models because it produced the highest accurately classified instances (77.78%) and the lowest inaccurately classified instances (22.22%). Although Cohen's kappa value is the highest value at 0.3047, NB only constructed a fair agreement that the model can separate the instances into the right class. PCA improves the performance of REPTree by 1.52%, increasing the correctly classified instances and 1.52%, reducing the inaccurately classified instances. Kappa statistic of 0.1927 shows the improvement of REPTree from an ineffective classifier to a slight agreement that it can be a promising breast cancer recurrence classifier.

Table 6. Summary statistic for classification models with PCA

Classifier	Accurately Classified Instances	Inaccurately Classified Instances	Kappa Statistic
NB	154 (77.78%)	44 (22.22%)	0.3047
REPTree	152 (76.77%)	46 (23.23%)	0.1927
KNN	135 (68.18%)	63 (31.82%)	0.1147

The confusion matrix analysis for the classification models with feature extraction (Table 7) verifies the results listed in Table 6. For each model, the sum of accurately classified instances equals the summation of TP and TN, and the sum of inaccurately classified instances equals the summation of FN and FP. For instance, the sum of accurately classified instances classified by NB is denoted as Accurately classified instances = TP+TN = 17+137 = 154.

Table 7. Confusion matrix for classification models with PCA

Classifier	R[a]	NR[b]	Classified As
Naïve Bayes	17 (TP)	30 (FN)	R
	14 (FP)	137 (TN)	NR
REPTree	10 (TP)	37 (FN)	R
	9 (FP)	142 (TN)	NR
KNN	15 (TP)	32 (FN)	R
	31 (FP)	120 (TN)	NR

[a.] Recurrence, [b.] Non-recurrence

Table 8 portrays the detailed accuracy by class for the classification models with feature extraction. The values of the F-measure (0.761) corroborate that NB is the best classification model that can be applied with PCA in predicting whether a patient has a breast cancer recurrence.

Table 8. Detailed accuracy for classification models with PCA

Classifier	P	R	F	Class
Naïve Bayes	0.548	0.362	0.436	R[a]
	0.820	0.907	0.862	NR[b]
Weighted Average	0.756	0.778	0.761	
REPTree	0.526	0.213	0.303	R
	0.793	0.940	0.861	NR
Weighted Average	0.730	0.768	0.728	
KNN	0.326	0.319	0.323	R
	0.789	0.795	0.792	NR
Weighted Average	0.679	0.682	0.681	

[a.] Recurrence, [b.] Non-recurrence

4.3. Comparative Analysis

The comparative analysis process entailed carrying out a comparison between models produced with and without PCA to determine the impact of decreasing feature dimensionality through principal component analysis on the outcome results. Fig. 4, Fig. 5, and Fig. 6 depict the performance for the three classification models built with and without PCA. The results show that when PCA is used for feature extraction, the performances of NB and REPTree improve by increasing the number of accurately classified instances, decreasing the number of inaccurately classified instances, and increasing the value of Cohen's kappa. However, this trend is not observed for KNN.

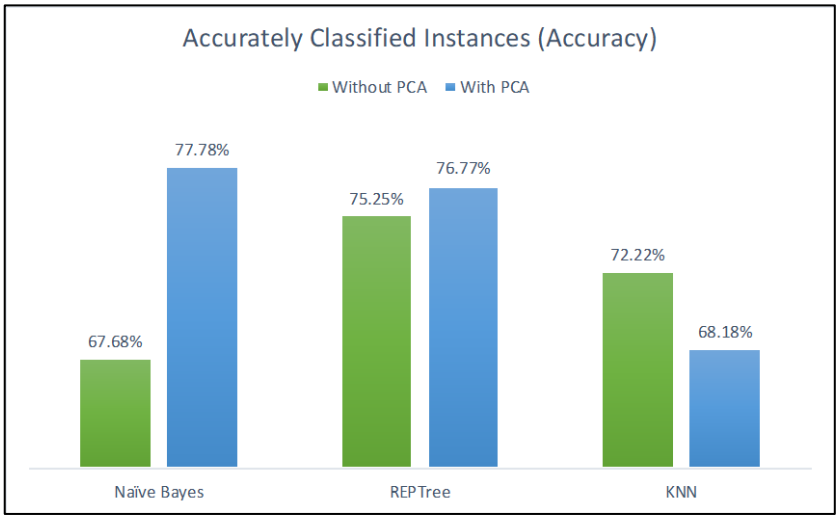


Fig. 4. Accuracy for three models with and without PCA

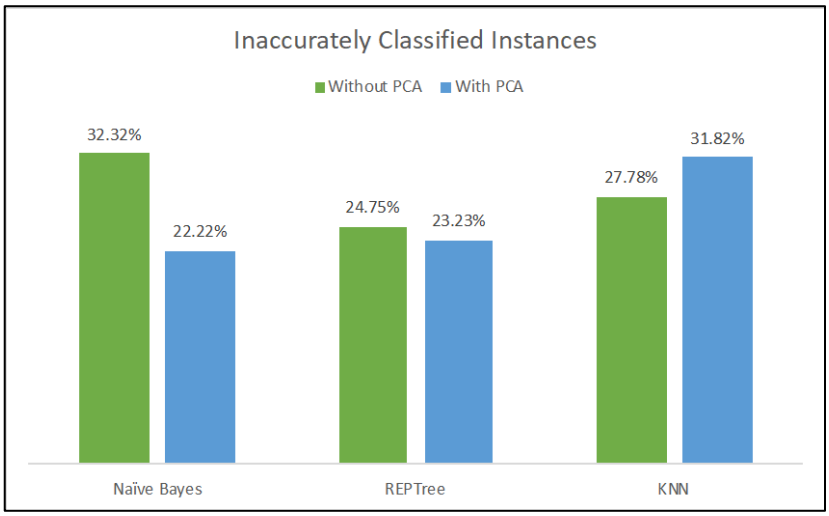


Fig. 5. Inaccurately classified instances for three models with and without PCA

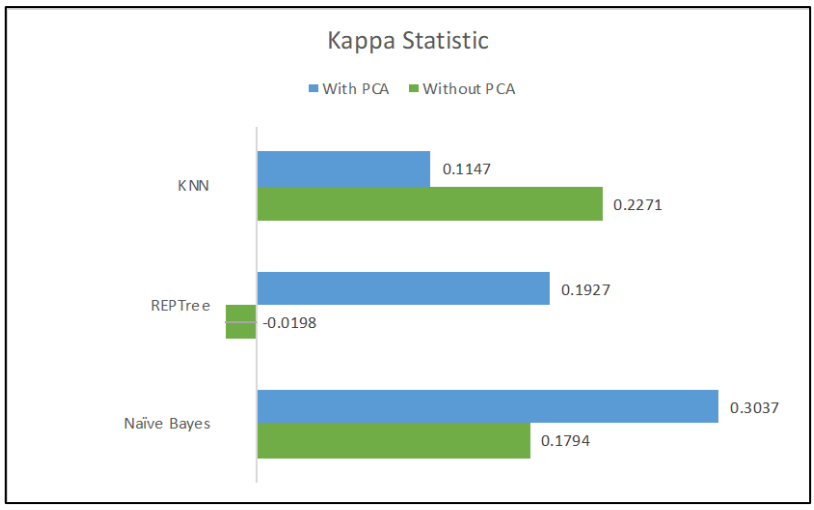
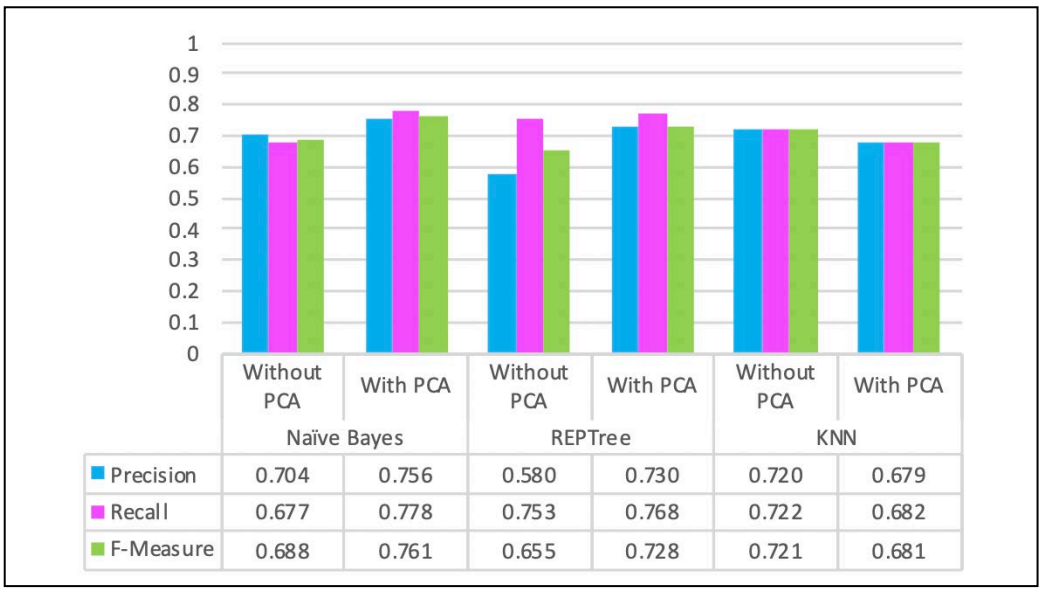


Fig. 6. Kappa statistic for three models with and without PCA

Fig. 7 presents the weighted average for each performance measure (precision, recall, and F-measure) for every classification model with and without PCA. The results confirm that the performances of NB and REPTree improve with PCA as feature extraction. This trend is not observed for KNN. It also

exposes that the classification model (NB built with PCA) is superior against the other five classification models. A much higher recall of the NB (77.8%) built with PCA denotes its exceptional potential in predicting the recurrence case out, which is specifically essential for actual breast cancer patients.



	Naïve Bayes		REPTree		KNN	
	Without PCA	With PCA	Without PCA	With PCA	Without PCA	With PCA
Precision	0.704	0.756	0.580	0.730	0.720	0.679
Recall	0.677	0.778	0.753	0.768	0.722	0.682
F-Measure	0.688	0.761	0.655	0.728	0.721	0.681

Fig. 7. Precision, recall, and F-measure for three models with and without PCA

4.4. The Difference from Prior Work

This study is unique and different from prior works in several ways. This research study was carried out by abiding by the key tenets of a systematic technique that amalgamated PCA with three popular classifiers, namely, KNN, NB, and REPTree, to forecast the recurrence associated with breast cancer, an entirely new experiment. Second, the comparative analysis was performed between the three classifiers, not only with PCA but also without PCA as a control to the experiment. Finally, this experiment's findings have been deliberated thoroughly by employing a raft of performance metrics, key among them being accurately classified instances, inaccurately classified instances, F-measure, Kappa statistics, recall, confusion matrix, and precision to avert bias.

5. Conclusion

This investigation aimed to compare and improve the performance of three established data mining algorithms, namely, NB, KNN, and REPTree, using PCA for feature extraction in predicting breast cancer recurrence. The comparison was conducted between models built with and without PCA. PCA, an unsupervised learning method, was employed to remove the repeated data and extract novel principal components to substitute the initial feature data. To carry out the study, a threshold of 95% was used to decrease the feature's dimension from 33 to 13 while retaining various principal components that signified roughly 95% variance between the initial dataset. These preprocessing stages provided a greatly valuable and reduced feature set that allows the MLA to train a classifier. The comparative analysis results revealed that PCA's involvement significantly improved the classifier's breast cancer recurrence detection ability for the WPBC dataset. Overall, this study strengthens the idea that without feature extraction, NB and REPTree's performance falls short in the ability to detect breast cancer recurrence. In contrast, applying PCA to cultivate and decrease the number of features increases the breast cancer recurrence detection possibility of NB by approximately 10% and REPTree by 2%, which is crucial to real patients with breast cancer. The results disclose the significance of minimizing feature dimensionality, particularly to classifiers whose performances can be significantly affected by the considerable quantity of features. In conclusion, this study shows that two out of three classifiers, NB and REPTree, outperformed when applying PCA as feature extraction with F-measure values equal to 76.1% and 72.8%, respectively. Thus, it can be considered to improve breast cancer recurrence prediction

of the WPBC dataset by researchers and practitioners. Further research should be carried out to explore another feature extraction technique in decreasing the dimensionality of the prognostic breast cancer set of data to improve classification models' performance in predicting recurrence. We should also study machine learning techniques to handle the imbalanced data issue in the prognostic breast cancer dataset.

Acknowledgment

The authors would like to thank the Information Systems Department, College of Computer & Information Sciences, Princess Nourah bint Abdulrahman University, KSA for providing facilities to conduct the research.

Declarations

Author contribution. All authors have equally contributed to this article. As well, all authors have read and approved the final version of the article.
Funding statement. The authors received no specific funding for this work.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] World Health Organization (WHO), "Breast cancer," 2020. [Online]. Available: www.who.int. [Accessed: 30-Oct-2020].

[2] H. Pan *et al.*, "20-Year Risks of Breast-Cancer Recurrence after Stopping Endocrine Therapy at 5 Years," *N. Engl. J. Med.*, vol. 377, no. 19, pp. 1836–1846, Nov. 2017, doi: 10.1056/NEJMoa1701830.

[3] A. Bhardwaj and A. Tiwari, "Breast cancer diagnosis using Genetically Optimized Neural Network model," *Expert Syst. Appl.*, vol. 42, no. 10, pp. 4611–4620, Jun. 2015, doi: 10.1016/j.eswa.2015.01.065.

[4] M. Seera and C. P. Lim, "A hybrid intelligent system for medical data classification," *Expert Syst. Appl.*, vol. 41, no. 5, pp. 2239–2249, Apr. 2014, doi: 10.1016/j.eswa.2013.09.022.

[5] W.-C. Yeh, W.-W. Chang, and Y. Y. Chung, "A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 8204–8211, May 2009, doi: 10.1016/j.eswa.2008.10.004.

[6] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1476–1482, Mar. 2014, doi: 10.1016/j.eswa.2013.08.044.

[7] I. M. D. Maysanjaya, I. M. A. Pradnyana, and I. M. Putrama, "Classification of breast cancer using Wrapper and Naïve Bayes algorithms," *J. Phys. Conf. Ser.*, vol. 1040, p. 012017, Jun. 2018, doi: 10.1088/1742-6596/1040/1/012017.

[8] L. Yang and Z. Xu, "Feature extraction by PCA and diagnosis of breast tumors using SVM with DE-based parameter tuning," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 3, pp. 591–601, Mar. 2019, doi: 10.1007/s13042-017-0741-1.

[9] S. A. Kumaraswamy and R. Mallika, "Cancer Classification in Microarray Data Using Gene Expression with KNN and FNN," *Int. J. Adv. Res. Comput. Sci.*, vol. 2, no. 5, 2011, doi: 10.26483/ijarcs.v2i5.722

[10] N. Sharma and H. Om, "Data mining models for predicting oral cancer survivability," *Netw. Model. Anal. Heal. Informatics Bioinforma.*, vol. 2, no. 4, pp. 285–295, Dec. 2013, doi: 10.1007/s13721-013-0045-7.

[11] J. Thongkam, G. Xu, Y. Zhang, and F. Huang, "Toward breast cancer survivability prediction models through improving training space," *Expert Syst. Appl.*, vol. 36, no. 10, pp. 12200–12209, Dec. 2009, doi: 10.1016/j.eswa.2009.04.067.

[12] C.-H. Jen, C.-C. Wang, B. C. Jiang, Y.-H. Chu, and M.-S. Chen, "Application of classification techniques on development an early-warning system for chronic illnesses," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 8852–8858, Aug. 2012, doi: 10.1016/j.eswa.2012.02.004.

[13] S. J, "Designing a Cloud Based Framework for Enhancing the Performance of Diabetic Classification Using Naïve Bayes Classifier," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 9, pp. 723–726, Sep. 2017, doi: 10.26483/ijarcs.v8i9.5204.

[14] J.-Y. Yeh, T.-H. Wu, and C.-W. Tsao, "Using data mining techniques to predict hospitalization of hemodialysis patients," *Decis. Support Syst.*, vol. 50, no. 2, pp. 439–448, Jan. 2011, doi: 10.1016/j.dss.2010.11.001.

[15] M. M. Kirmani and S. I. Ansarullah, "Classification models on cardiovascular disease detection using Neural Networks, Naïve Bayes and J48 Data Mining Techniques.," *Int. J. Adv. Res. Comput. Sci.*, vol. 7, no. 5, 2016, Available at: Google Scholar

[16] S. Fei, "Diagnostic study on arrhythmia cordis based on particle swarm optimization-based support vector machine," *Expert Syst. Appl.*, vol. 37, no. 10, pp. 6748–6752, Oct. 2010, doi: 10.1016/j.eswa.2010.02.126.

[17] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015, doi: 10.1016/j.csbj.2014.11.005.

[18] N. M. P. Trushna Patel, Darshak G Thakore, "A Survey on Object Detection Based Automatic Image Captioning using Deep Learning," *Int. J. Mod. Trends Sci. Technol.*, vol. 6, no. 4, pp. 274–280, 2020, Available at: ijmtst.com

[19] A. Jamal, A. Handayani, A. A. Septiandri, E. Ripmiatin, and Y. Effendi, "Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction," *Lontar Komput. J. Ilm. Teknol. Inf.*, p. 192, Dec. 2018, doi: 10.24843/LKJITI.2018.v09.i03.p08.

[20] J. Verma, M. Nath, P. Tripathi, and K. K. Saini, "Analysis and identification of kidney stone using Kth nearest neighbour (KNN) and support vector machine (SVM) classification techniques," *Pattern Recognit. Image Anal.*, vol. 27, no. 3, pp. 574–580, Jul. 2017, doi: 10.1134/S1054661817030294.

[21] W. N. H. W. Mohamed, M. N. M. Salleh, and A. H. Omar, "A comparative study of Reduced Error Pruning method in decision tree algorithms," in *2012 IEEE International Conference on Control System, Computing and Engineering*, 2012, pp. 392–397, doi: 10.1109/ICCSCE.2012.6487177.

[22] C. J. C. Burges, "Dimension Reduction: A Guided Tour," *Found. Trends® Mach. Learn.*, vol. 2, no. 4, pp. 275–364, 2009, doi: 10.1561/2200000002.

[23] Tsang-Hsiang Cheng, Chih-Ping Wei, and V. S. Tseng, "Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches," in *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, 2006, pp. 165–170, doi: 10.1109/CBMS.2006.87.

[24] G. Pfurtscheller *et al.*, "Graz-BCI: state of the art and clinical applications," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 11, no. 2, pp. 1–4, Jun. 2003, doi: 10.1109/TNSRE.2003.814454.

[25] H. Hasan and N. M. Tahir, "Feature selection of breast cancer based on Principal Component Analysis," in *2010 6th International Colloquium on Signal Processing & its Applications*, 2010, pp. 1–4, doi: 10.1109/CSPA.2010.5545298.

[26] S. Jhajharia, H. K. Varshney, S. Verma, and R. Kumar, "A neural network based breast cancer prognosis model with PCA processed features," in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2016, pp. 1896–1901, doi: 10.1109/ICACCI.2016.7732327.

[27] M. S. Uzer, O. Inan, and N. Yılmaz, "A hybrid breast cancer detection system via neural network and feature selection based on SBS, SFS and PCA," *Neural Comput. Appl.*, vol. 23, no. 3–4, pp. 719–728, Sep. 2013, doi: 10.1007/s00521-012-0982-6.

[28] K. Bian, M. Zhou, F. Hu, and W. Lai, "RF-PCA: A New Solution for Rapid Identification of Breast Cancer Categorical Data Based on Attribute Selection and Feature Extraction," *Front. Genet.*, vol. 11, Sep. 2020, doi: 10.3389/fgene.2020.566057.

[29] R. H and A. T, "Feature Extraction of Chest X-ray Images and Analysis using PCA and kPCA," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 5, p. 3392, Oct. 2018, doi: 10.11591/ijece.v8i5.pp3392-3398.

[30] S. Ray, "6 Easy Steps to Learn Naive Bayes Algorithm," 2017. [Online]. Available: analyticsvidhya [Accessed: 04-Jan-2020].

[31] F. Provost and R. Kohavi, "Guest editors' introduction: On applied research in machine learning," *Mach. Learn.*, vol. 30, no. 2–3, pp. 127–132, 1998, doi: 10.1023/A:1007442505281.

[32] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, "Data Mining: Practical Machine Learning Tools and Techniques," in *Data Mining*, Elsevier, 2017, pp. 417–466, doi: 10.1016/B978-0-12-804291-5.00010-6

[33] J. R. Quinlan, "Simplifying decision trees," *Int. J. Man. Mach. Stud.*, vol. 27, no. 3, pp. 221–234, Sep. 1987, doi: 10.1016/S0020-7373(87)80053-6.

[34] "University of California Irvine Machine Learning Repository." [Online]. Available: ics.uci.edu. [Accessed: 01-Aug-2019].

[35] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977, doi: 10.2307/2529310.