IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 5, 2018, accepted May 18, 2018, date of publication June 4, 2018, date of current version June 20, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2843443

Particle Swarm Optimization Feature Selection for Breast Cancer Recurrence Prediction

SAPIAH BINTI SAKRI, NURAINI BINTI ABDUL RASHID, AND ZUHAIRA MUHAMMAD ZAIN

College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia

Corresponding author: Sapiah Binti Sakri (sbsakri@pnu.edu.sa)

This work was supported by the Deanship of Scientific Research at Princess Nourah bint Abdulrahman under Grant ٣٨-ص-128.

ABSTRACT Women who have recovered from breast cancer (BC) always fear its recurrence. The fact that they have endured the painstaking treatment makes recurrence their greatest fear. However, with current advancements in technology, early recurrence prediction can help patients receive treatment earlier. The availability of extensive data and advanced methods make accurate and fast prediction possible. This research aims to compare the accuracy of a few existing data mining algorithms in predicting BC recurrence. It embeds a particle swarm optimization as feature selection into three renowned classifiers, namely, naive Bayes, K-nearest neighbor, and fast decision tree learner, with the objective of increasing the accuracy level of the prediction model.

INDEX TERMS Breast cancer, recurrence, feature selection, REPTree, naïve Bayes, K-nearest neighbor, particle swarm optimization.

I. INTRODUCTION

Today, cancer is the primary cause of death around the globe [1]. As stated by Seigel, breast cancer (BC) will continue to be the most prevalent cancer in women [1]. Every woman is at risk for breast cancer. If she is 85 years old, there is a one in eight chance (12%) that she will develop breast cancer once during her life [2].

In 2010, breast cancer was ranked the ninth leading cause of death in the Kingdom of Saudi Arabia (KSA) [3], [4]. Before that, in 2009, it was reported that there were 1,308 new breast cancer cases, representing 25% of registered new cancer cases among Saudi women [5]. It was forecasted that this disease incidence would increase over the coming decades in KSA as the population grows and ages [6]. It is also notable that obesity; having a first child at a late age, a young age at menarche, a short period of lactation, or an unhealthy lifestyle; and geographical, racial and ethnic characteristics are risk factors contributing to the cause of breast cancer [3], [7]–[11].

Previous research [12]–[14] indicated that the characteristics of this disease are high aggressiveness, poor clinicopathologic features, and early onset among the Saudis. Studies also reported that the advanced stage of breast cancer disease was found to be more prevalent in younger women with a median age of 47 years than in older women with a median age of 63 years in industrialized nations [5], [14]–[15].

From the perspective of breast cancer behavior, recurrence of breast cancer refers to the reoccurrence of breast cancer in a patient whose previous cancer had gone into remission. Remission is the desired result of chemotherapy and continual treatment by oncologists. Recurrence of breast cancer or any other cancer is among the most significant fears faced by a cancer patient. Consequently, it becomes one of the concerns that affect their quality of life. Regardless of its relevance, it is infrequently recorded in most breast cancer datasets, which makes research into its prediction more problematic. In addition to the obvious mortality ramifications of recurrence, BC patients also confront severe treatment-related intricacies, which increases their risk of death from causes irrelevant to breast cancer itself [16]. Accurate prediction of BC behavior assumes an essential role in this situation, as it helps clinicians in their decision-making process, supporting a more personalized treatment for patients.

Methods such as knowledge discovery in databases (KDD) provide an exciting avenue to investigate such data driven problems. Data mining, a subset of KDD, is an iterative process in the search for new, valuable, and non-trivial information in large volumes of data [17]. The data mining and machine learning approaches have been successfully used in diagnosing and predicting various health-related diseases. These include breast cancer [2], [18]–[24], oral cancer [25], cardiovascular diseases [26]–[29], and diabetes [26], [30].

The results from these successful studies are used as a motivator to apply data mining technologies as a predictive tool for breast cancer recurrence prediction. Thus, utilizing data mining in these specific forms is the basis of this research.

With the advancement of high-throughput technologies, various types of high-dimensional data have been generated in recent years, specifically those related to disease occurrence or management of cancer recurrence. The high dimensionality of the data makes it more difficult to obtain insights from them. There is an urgent need to convert high dimensional data to low dimensional data by using dimensionality reduction methods. Dimensionality reduction facilitates the classification, visualization, communication, and storage of high dimensional data.

The dimensions of medical data include several features, and each consists of different types of values. Data quality-related issues include the presence of noise, outliers, missing or duplicate data, and data that are biased-unrepresentative [31]. Preprocessing steps should be applied to make the raw data more suitable for further analysis, focusing on the data preparation. The feature selection method can be used to overcome some of these problems. Feature selection is an essential step in building the classifier; it is even said that limiting the number of input features in a classifier is advantageous to creating an excellent predictive and less computationally intensive model. This distinction is vital since, in the area of medical diagnosis, a small feature subset means lower test and diagnostic costs [23]. Additionally, the reduction of dimensionality can eliminate irrelevant features, while the reduction in noise can produce more robust learning models due to the association of fewer features [31]. Most of the traditional feature selection methods select the most relevant, non-redundant features but disregard the fundamental interdependent structure of the features.

This study proposed the particle swarm optimization (PSO) algorithm as the feature selection method in reducing the high dimensionality of the Wisconsin Prognosis Breast Cancer dataset, with three renowned classification algorithms as the classifiers, in an effort to analyze the accuracy level of these three different prediction models. The algorithms are the naive Bayes, K-nearest neighbor (IBK), and fast decision tree learner (REPTree). We also conducted a comparative analysis of the performance metrics between the original dataset and the dataset that has selected features or attributes only.

The remainder of this paper contains Section II, which corresponds to a review of all related works within the study domain. The review includes the background information on breast cancer research, prognosis factors, uses of ranking algorithms, several data mining techniques for breast cancer estimation, and a comparison of their accuracies. Section III describes the information about the Wisconsin Prognosis Breast Cancer Dataset that was used to experiment with the three algorithms and various other testing processes. Section IV explains the performance evaluation method. Section V then discusses the experimental results of this study, mainly focusing on the performance evaluation per the aim of the study. Finally, Section VI presents the conclusions of the study and highlights the scope of future work.

II. RELATED WORK

A. DIMENSIONALITY ISSUES

Dimension reduction is defined by Burgess as the mapping of data to a lower dimensional space by removing uninformative variance in data such that a subspace in which the data reside is then detected [32]. Dimension reduction can be divided into feature extraction and feature selection [32]. Feature extraction is the process of distinguishing and disregarding irrelevant, less relevant, or redundant attributes of dimensions in a given dataset [32], [33]. With feature selection, it is possible to identify and remove as much irrelevant and redundant information as possible to build robust learning models [32], [33]. Thus, feature selection not only reduces the computational and processing costs but also improves the model developed from the selected data [34], [35]. A number of existing works have been performed using the feature selection method on healthcare data [36]–[39]. The feature selection methods can be categorized into three types of algorithms: filters, wrappers, and embedded approaches [40].

Dimension reduction, as explained by Burgess, is the mapping of data to a lower dimensional space such that uninformative variance in the data is removed such that a subspace in which the data reside is detected [32]. We can further divide dimension reduction into instance selection or reduction and feature selection techniques [32]. Instance reduction is the process of reducing the irrelevant instances from the dataset to increase the classification accuracy, while feature selection is the selection of a subset of the relevant features used in the model construction [32]. These irrelevant instances are not beneficial for classification and may reduce the classification performance. Feature selection helps in removing irrelevant, redundant, and noisy features that are not instrumental to the accuracy of the model [33]. Therefore, it becomes easier to determine only the useful and relevant features for classification rather than using all of them [39]. This results in a fewer number of features, which is desirable, as it simplifies the model and makes it easier to understand. Implementing feature selection in healthcare data will reduce the number of tests required to identify a disease, saving time and money for the patient undergoing tests. In general, we can broadly divide traditional feature selection algorithms into three classes, which are filter approaches, wrapper approaches, and embedded approaches [39]–[42].

B. DATA MINING IN HEALTHCARE

Every day, the size of data is increasing; therefore, the need to understand large and complex data is also growing in varied fields, including business, medicine, science, and many others. The ability to extract useful information hidden in this vast amount of data and act on the information is becoming an increasingly important challenge in today's competitive world.

The data mining standard is grounded in disciplines such as machine learning, artificial intelligence, probability, and statistics [43]. There are two kinds of data mining models: predictive models and descriptive models [43]. A predictive model is usually applied to supervised learning functions to predict unknown or future values of the variables of interest [43]. Meanwhile, unsupervised learning functions use a descriptive model in finding patterns to describe the data that can be interpreted by humans [43].

To implement a model, we need to use a data mining task. The tasks to implement descriptive models are clustering [44], association rules [45], correlation analysis [46], and anomaly detection [47]. For predictive models, the tasks often used are classification [48], regression [49], and categorization [50]. Once the data mining model and task are defined, the suitable data mining methods will be used to build the approach based on the discipline. The methods often used for anomaly detection are standard support vector data description, density-induced support vector data description, and Gaussian mixture [47]. The vector quantization method is used for the clustering task [51].

The methods used for classification in healthcare are statistical methods [52], discriminant analysis [53], decision trees [54], Markov-based methods [55], swarm intelligence [19], k-nearest neighbor [56], genetic classifiers [57], artificial neural networks [58], support vector [23], and association rules [59].

From the literature reviewed, there are various data mining systems designed for healthcare industry use; however, there is still lack of research on predicting breast cancer recurrence. Therefore, this study was designed to close this gap by combining the feature selection and classification algorithms to predict breast cancer recurrence. The combined algorithm was designed to reduce the training and utilization times and to overcome the curse of high dimensional data among the breast cancer parameters, using a filter model to increase the prediction accuracy.

III. METHODOLOGY

The overall research methodology for this study was adapted based on the knowledge discovery process and is illustrated in Figure 1. The data acquisition phase was the first phase of this methodology, in which we obtained the relevant data for the study. The second phase was the data preprocessing stage, in which the collected information was integrated, cleaned, and transformed such that the datasets were suitable for classification prediction. After this, still in the second phase, we carried out feature extraction. The data from the preprocessing stage (Phase 2) were then carried over to Phase 3 for classification prediction. In Phase 4, the enhanced prediction algorithm, based on the particle swarm, was designed, trained, and tested on the data for classification prediction. In the final phase of the research, we performed a comparative analysis of the models without feature selection and the models that used feature selection.

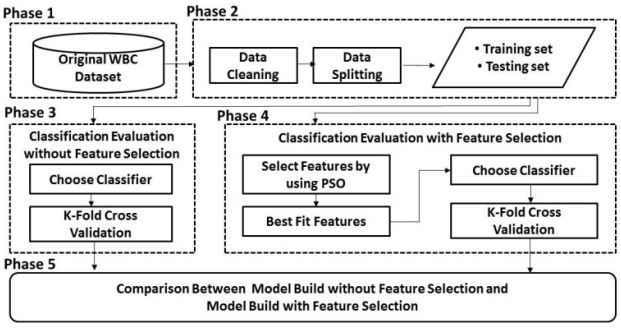

FIGURE 1. Research Methodology.

A. PHASE 1 – DATA ACQUISITION

In the data selection phase, we collected breast cancer data from the UCI public database [60]. The Breast Cancer Wisconsin Breast Cancer Prognostic Dataset has 198 instances and 34 attributes.

B. PHASE 2 – DATA PREPROCESSING

1) DATA CLEANING

The integrated database went through the data cleaning process, in which we removed improper data entries, such as those that provided an irrelevant answer, in the database. To smooth noisy data, the tuples with improper data entry were eliminated or filled with the most probable value, as this is one of the most popular strategies to counter this issue. Additionally, the find and replace function was used to handle inconsistency in the format of data from the survey.

2) DATA SPLITTING

In the data splitting phase, we divided the data into two datasets: the training dataset and the test dataset. The standard proportion of the splitting process is 60% training and 40% testing [61]. The purpose of splitting the data is to ensure that the model is not overfitted during the model testing with the testing dataset.

C. PHASE 3 – CLASSIFICATION EVALUATION WITHOUT FEATURE SELECTION

In this phase, we constructed three models by using three renowned algorithms, namely, naïve Bayes, REPTree, and KNN (IBK), as the classifier with the test option of 10-fold cross-validation. The training dataset with all the features/attributes was used for the evaluation.

D. PHASE 4 – CLASSIFICATION EVALUATION WITH FEATURE SELECTION

In this phase, the process of feature selection was performed by using PSO in an effort to acquire the best-fit features. Then, we used the selected features to build the three models by performing the same process described in the above paragraph.

E. PHASE 5 – COMPARATIVE ANALYSIS

In this phase, we compared the three models without PSO feature selection and the three models with PSO feature selection. Before performing this analysis, we tested all the models for fitness. The method of testing the fitness of the prediction model was examining the confusion matrix [62], [23]; the confusion matrix [63] contained information about the actual and predicted classification obtained by the proposed classifier. The proposed model was validated and benchmarked with the help of an oncologist to evaluate the classification and verify the correctness of the prediction model. The other measures assessed for effectiveness were classification accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and ROC curves [62], [23]. Figure 2 and Figure 3 show the confusion matrix and classification measure representations, respectively. The proposed method was evaluated against existing data mining methods for accuracy, correctness, and effectiveness.

Confusion Matrix Representation		
Actual	Predicted	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

FIGURE 2. Confusion Matrix Representation.

$$\text{Classification Accuracy (\%)} = \frac{TP + TN}{TP + FP + FN + TN}$$

$$\text{Sensitivity (\%)} = \frac{TP}{TP + FN}$$

$$\text{Specificity (\%)} = \frac{TN}{FP + TN}$$

$$\text{Positive predictive value} = \frac{TP}{TP + FP}$$

$$\text{Negative predictive value} = \frac{TN}{FN + TN}$$

FIGURE 3. Classification Measure Representation.

In this phase, we compared the results of the prediction models constructed without using PSO feature selection to those constructed with the use of PSO feature selection. The comparative analysis was conducted in terms of model correctness, statistical details, standard errors, and precision of the predicting models.

IV. DATASET DESCRIPTION

For this research, which focused on the methods and techniques previously discussed, the study leveraged the available dataset provided by the UC Irvine machine learning repository, acquired from the Wisconsin Prognostic Breast Cancer sub-directory with 198 instances [60]. A description of the dataset's attributes and domains is provided in Table 1.

TABLE 1. Attribute information of the dataset.

No.	Attribute	Domain
1.	ID code number	Id Number
2.	Outcome	R = Recur, N = Non-recur
3.	Time	1 – 4 (recurrence time if field 2= R, disease-free time if field 2= N)
4.	Features Value #1	1 - 4
:	: : : : :	: :
33.	Features Value #32	1 - 10
34.	Tumor size	The diameter of the excised tumor in centimeters
35.	Lymph node status	The number of positive axillary lymph nodes observed at the time of surgery

The first 30 features were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass sample. They describe characteristics of the cell nuclei present in the image.

Ten real-valued features were computed for each cell nucleus:

- radius (mean of distances from the center to points on the perimeter)
- texture (standard deviation of gray scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension (''coastline approximation'' - 1)

For each image, the mean, standard error, and ''worst'' or largest (the mean of the three largest values) of these features were computed, resulting in 30 features. For instance, field 4 is Mean Radius, field 14 is Radius SE, and field 24 is Worst Radius. Values for features 4-33 were re-coded with four significant digits. There were four cases of missing values for the Lymph node status attribute. The class distribution was found to include 151 non-recurs and 47 recurs. Each record represents follow-up data for one breast cancer case. These are consecutive patients treated by Dr. Wolberg since 1984 [64] and comprise only those cases presenting invasive breast cancer and no evidence of distant metastases at the time of diagnosis.

The other attributes are listed in Table 1, which summarizes all 35 attributes of the dataset.

V. CLASSIFICATION ALGORITHM DESCRIPTIONS

In this study, three renowned [61] classification algorithms for the prediction model, namely, naive Bayes, fast decision tree learner, and K-nearest neighbor, were evaluated in the prediction of breast cancer recurrence by using the Wisconsin Prognostic Breast Cancer Dataset. The following paragraph briefly describes each of the algorithms.

A. NAIVE BAYES ALGORITHM

Bayesian classification represents a supervised learning method as well as a statistical method for classification. It assumes an underlying probabilistic model and allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. This classification is named after Thomas Bayes (1702-1761), who proposed the Bayes theorem. Bayesian classification provides practical learning algorithms, in which prior knowledge and observed data can be combined. Bayesian classification presents a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for a hypothesis, and it is robust to noise in input data.

B. FAST DECISION TREE LEARNER (REPTREE) ALGORITHM

The reduced error pruning (REP) tree classifier is a quick ''decision tree learning algorithm and is based on the principle of computing the information gain with entropy and minimizing the error arising from variance'' [65]. This algorithm was first recommended in [66]. REPTree applies regression tree logic and generates multiple trees in altered iterations. Afterward, it selects the best tree from all spawned trees. This algorithm constructs the regression/decision tree using variance and information gain. Additionally, this algorithm prunes the tree with reduced-error pruning using a backfitting method. It sorts the values of numeric attributes once at the beginning of the model preparation. Additionally, as in the C4.5 Algorithm, this algorithm also addresses missing values by splitting the corresponding instances into pieces [67].

C. K-NEAREST NEIGHBORS (IBK) ALGORITHM

K-nearest neighbors (KNN) is a supervised classification algorithm in which the $k$nearest neighbors of a point are chosen, found by minimizing a similarity measure (e.g., the Euclidean distance or Mahalanobis distance) [68]. To determine the class of an unlabeled example, KNN computes its distance to the remaining (labeled) examples and determines its k-nearest neighbors and respective labels. The unlabeled object is then classified either by majority voting—the dominant class in the neighborhood—or by a weighted majority, where greater weight is given to points closer to the unlabeled object.

VI. FEATURE SELECTION ALGORITHM DESCRIPTIONS

Classification involves conscientious consideration of the dataset before assigning the data to a classifier. The recommendation is to consider only necessary features to make the classification process much easier, rather than adding many irrelevant features. Therefore, it is beneficial to have sufficient techniques that are capable of selecting the relevant and significant features. Moreover, if feature selection is adopted in classification, it helps in finding the significant feature and reducing the workload of the classifier, which also improves the classification accuracy. Based on the review of the existing literature [69]–[75], particle swarm optimization enjoys better selection, in terms of classification accuracy, compared to other existing feature selection techniques. From the review [75]–[83], potential advantages of PSO for feature selection are as follows:

- PSO has a powerful exploration ability until the optimal solution is found because different particles can explore different parts of the solution space.
- PSO is particularly attractive for feature selection since the particle swarm has memory, and knowledge of the solution is retained by all particles as they fly within the problem space.
- The attractiveness of PSO is also due to its computationally inexpensive implementation that still gives decent performance.
- PSO works with the population of potential solutions rather than with a single solution.
- PSO can address binary and discrete data.
- PSO has better performance compared to other feature selection techniques in terms of memory and runtime and does not need complex mathematical operators.
- PSO is easy to implement with few parameters, is easy to realize and gives promising results.
- The performance of PSO is almost unaffected by the dimension of the problem.

Particle swarm optimization is a computation technique inspired by the simulation of social behavior, proposed by [82]. The original concept of PSO is to simulate the behavior of flying birds and their means of information exchange to solve problems. An overview to provide insight into how PSO works in the search for the significant features is given in Figure 4 [75].

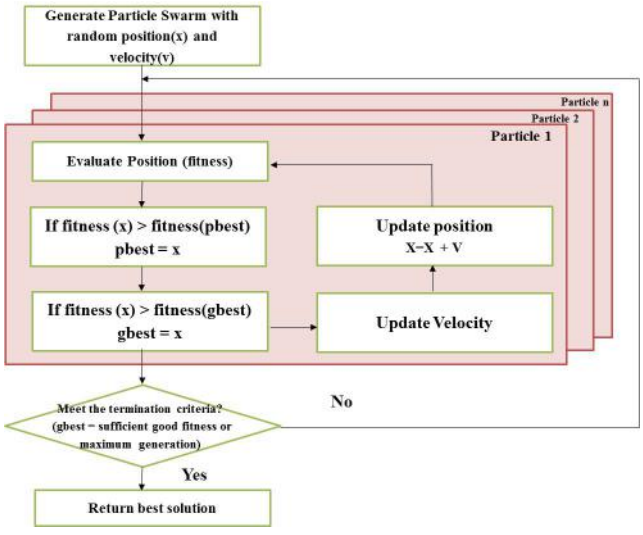

FIGURE 4. Particle Swarm Optimization (PSO) Process.

VII. EXPERIMENTAL RESULTS AND DISCUSSION

This section briefly describes the experimental results obtained in the three phases, namely, the classification

evaluation without feature selection phase, classification evaluation with feature selection phase, and comparative analysis phase.

A. CLASSIFICATION EVALUATION WITHOUT FEATURE SELECTION PHASE

In this phase, we constructed three prediction models using the training dataset for the naive Bayes, REPTree, and K-nearest neighbor classifiers. Ten-fold cross-validation was used as the test option for the models. The models were constructed by using all the attributes contained in the WBC prognostic dataset without any process of feature selection.

The results of the experiment, depicted in Figure 5, show the accuracy, specificity, and sensitivity rate of the prediction models for the naive Bayes classifier, REPTree classifier, and K-nearest neighbor classifier. Table 2 shows the detailed statistics for all the models. Table 3 shows the performance measure for all the models.

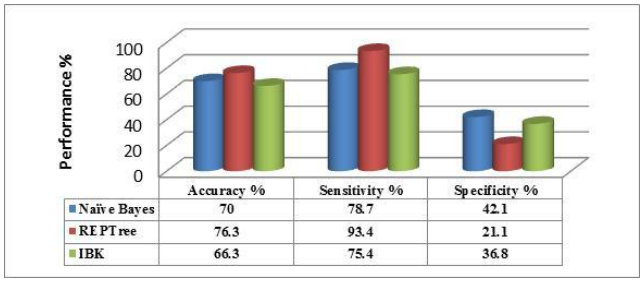

FIGURE 5. Performance of the Prediction Models without Feature Selection.

TABLE 2. Detailed statistics for all the models without feature selection.

	Naive Bayes	REPTree	IBK
Correctly Classified Instances	0.7	0.763	0.663
Incorrectly Classified Instances	0.3	0.238	0.338
Kappa Statistic	0.201	0.181	0.116
Mean Absolute Error	0.307	0.322	0.342
Root Mean Squared Error	0.535	0.436	0.573
Relative Absolute Error	0.838	0.879	0.934

The experiment results indicated that the accuracy percentage of the breast cancer recurrence prediction model that used REPTree as the classifier was higher compared to the naive Bayes and IBK classifiers. REPTree gave 76.3%, Naive Bayes gave 70% and IBK gave 66.3% accuracies. The dataset used in this experiment did not apply PSO feature selection.

B. CLASSIFICATION EVALUATION WITH FEATURE SELECTION PHASE

In this phase, two stages were involved. The first stage was to perform feature selection based on the requirements stated in Table 4. The second stage was to evaluate the performance

TABLE 3. Performance measures for all the models without feature selection.

	Precision	Recall	F Score	ROC Area	Class
Naive Bayes	0.814	0.787	0.800	0.758	Non-Recurrence
	0.381	0.421	0.400	0.758	Recurrence
REPTree	0.765	0.978	0.859	0.540	Non-Recurrence
	0.333	0.036	0.065	0.540	Recurrence
IBK	0.793	0.754	0.773	0.561	Non-Recurrence
	0.318	0.368	0.341	0.561	Recurrence

TABLE 4. PSO feature selection requirement.

Process	Requirement
Attribute Evaluator:	CFS Subset Evaluator (supervised, Class (nominal): 1 Outcome): Including a locally predictive attribute
Search Method:	PSOSearch
Requirement to obtain the best fit features	Start set: no attributes Population size: 20 Number of iterations: 20 Mutation type: bit-flip Mutation probability: 0.01 Inertia weight: 0.33 Social weight: 0.33 Individual weight: 0.34 Report frequency: 20 Seed: 1
Result:	
The best-selected features/attributes:	11,22,32,34 : 4 CN10 CN21 CN31 Time

of the models based on the selected features. The process of feature selection was executed by using the WEKA tool, which is the PSOSearch that explores the attribute space using the particle swarm optimization (PSO) algorithm.

The outcome of this phase was obtaining the best-fit features selected from the dataset. The results showed that out of 34 attributes, four attributes were found to be the best fit features - Cell Nucleus 10, Cell Nucleus 21, Cell Nucleus 31, and Time. Cell Nucleus was computed based on ten real-valued features, namely, the radius (mean of distances from the center to points on the perimeter), texture (standard deviation of gray scale values), perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter$\wedge$2 / area - 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, and fractal dimension (''coastline approximation'' - 1).

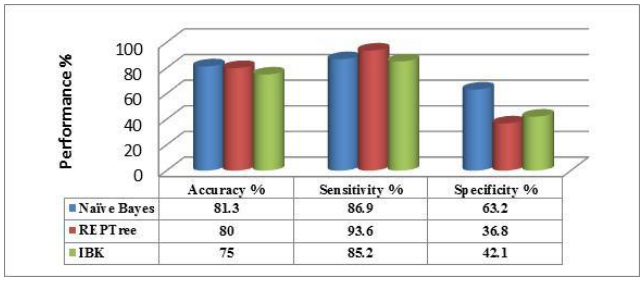

FIGURE 6. Performance of the Prediction Models with Feature Selection.

TABLE 5. Detailed statistics for all the models with feature selection.

	Naive Bayes	REPTree	IBK
Correctly Classified Instances	0.813	0.8	0.75
Incorrectly Classified Instances	0.188	0.2	0.25
Kappa Statistic	0.492	0.354	0.284
Mean Absolute Error	0.131	0.301	0.257
Root Mean Squared Error	0.381	0.401	0.493
Relative Absolute Error	0.646	0.823	0.701

The results of the experiment in terms of the percentage of accuracy among the classifiers are shown in Figure 6. The results indicated that naive Bayes gave 81.3% accuracy, the highest among the models. Table 5 shows the detailed statistics for all the models. Table 6 shows the performance measure for all the models.

TABLE 6. Performance measures for all the models with feature selection.

	Precision	Recall	F Score	ROC Area	Class
Naive Bayes	0.883	0.869	0.876	0.820	Non-Recurrence
	0.600	0.632	0.615	0.820	Recurrence
REPTree	0.826	0.934	0.877	0.672	Non-Recurrence
	0.636	0.368	0.467	0.672	Recurrence
IBK	0.825	0.852	0.839	0.637	Non-Recurrence
	0.471	0.421	0.444	0.637	Recurrence

From this evaluation, the results indicated that the prediction model that used naive Bayes as the classifier exhibited a high precision in predicting non-recurrence of breast cancer - 0.883. On the other hand, the prediction model using the REPTree classifier had the highest precision value of predicting breast cancer recurrence (0.636) as when compared to the other two prediction models using naive Bayes and IBK as the classifiers.

C. COMPARATIVE ANALYSIS PHASE

This phase involved discussion of the comparison results of the performance evaluation criteria or metrics based on the two experiments performed in the previous phases. Table 7 presents the comparative analysis. In this analysis, evaluation of the three performance metrics, accuracy, specificity, and sensitivity, was performed.

TABLE 7. Performance of classifiers.

Evaluation Criteria	Classifiers					
	Naive Bayes		REPTree		IBK	
	Without Feature Selection (PSO)	With Feature Selection (PSO)	Without Feature Selection (PSO)	With Feature Selection (PSO)	Without Feature Selection (PSO)	With Feature Selection (PSO)
Accuracy (%)	70.0%	81.3%	76.3%	80.0%	66.3%	75.0%
Sensitivity (%)	78.7%	86.9%	93.4%	93.6%	75.4%	85.2%
Specificity (%)	42.1%	63.2%	21.1%	36.8%	36.8%	42.1%

1) COMPARATIVE ANALYSIS OF THE THREE PERFORMANCE METRICS

In this research, we used three performance metrics, namely, accuracy, specificity, and sensitivity. Hence, the results from the two experiments were reported based on the three metrics, as shown in Table 7.

The results show that when using PSO feature selection, all the classifiers out-performed their counterparts without feature selection in terms of the accuracy level. The best result of the accuracy level for breast cancer recurrence was exhibited by the naive Bayes classifier, which obtained 81.3% accuracy, followed by REPTree, with 80% accuracy, and K-nearest neighbor (IBK) reaching the 75.0% accuracy level.

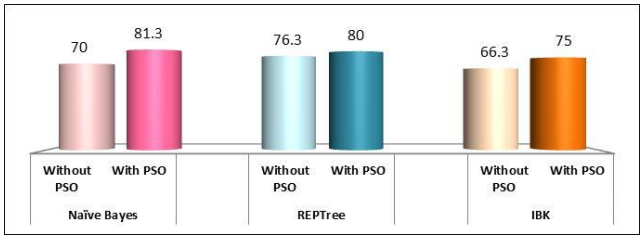

FIGURE 7. Accuracy (%) of Classification Algorithms with and without PSO Feature Selection.

However, in the experiment without PSO feature selection, REPTree classifier provided the highest accuracy level of 76.3%, followed by naive Bayes at 70% and IBK at 66.3%. Graphical representations of the results are shown in Figure 7, Figure 8, and Figure 9. The results indicated that proper attribute selection for classification and the reduction of high dimensionality of the dataset could improve the results by a fair margin. A graphical representation of the overall results of the experiment regarding the performance metrics is as shown in Figure 10.

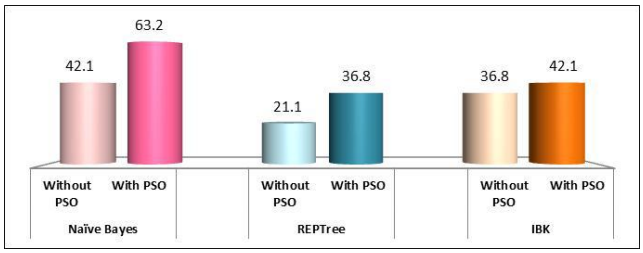

FIGURE 8. Specificity (%) of Classification Algorithms with and without PSO Feature Selection.

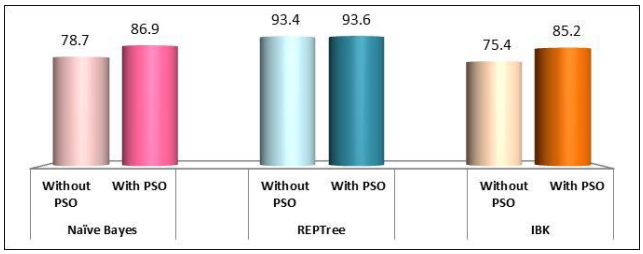

FIGURE 9. Sensitivity (%) of Classification Algorithms with and without PSO Feature Selection.

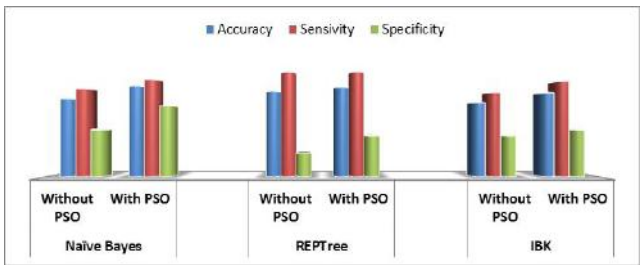

FIGURE 10. Performance Metrics of All the Classification Algorithms with and without PSO Feature Selection.

2) COMPARATIVE ANALYSIS BASED ON THE DETAILED STATISTICAL RESULTS OF EACH CLASSIFIER

The next performance evaluation criteria pertained to the detailed statistical results of each of the classifiers. Again, the two experiments were performed to enable the comparative analysis between the dataset with the original features and the dataset with the PSO selected feature. The graphical representations of the results are shown in Figure 11, Figure 12, Figure 13, and Figure 14.

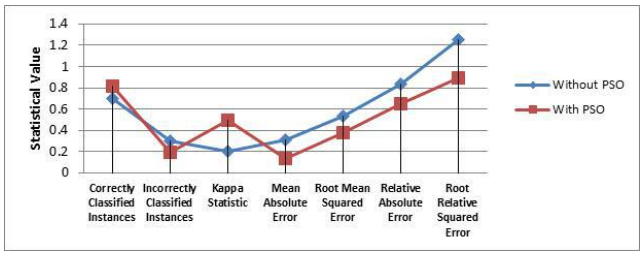

FIGURE 11. Evaluation Results for the Naïve Bayes Classifier.

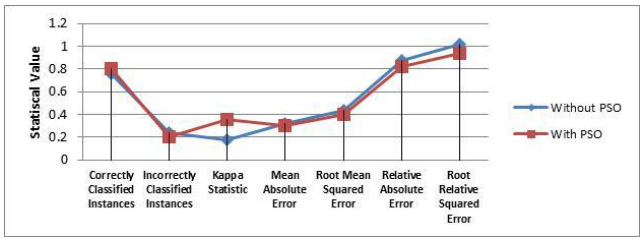

FIGURE 12. Evaluation Results for the REPTree Classifier.

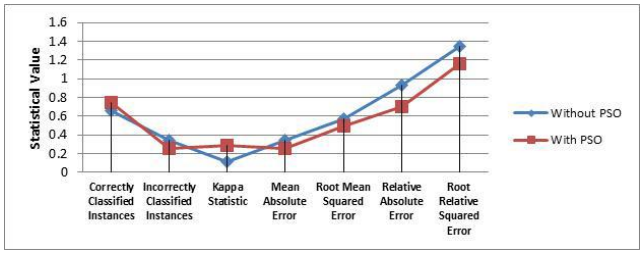

FIGURE 13. Evaluation Results for the K-Nearest Neighbor Classifier.

FIGURE 14. Comparison between All the Classifiers.

The accuracy levels, as gleaned from the correctly classified instances, had undoubtedly improved when the models used PSO feature selection. The error rate for every evaluation criteria was reduced significantly after including the PSO feature selection algorithm. The Kappa statistic for all classifiers also improved when PSO feature selection was embedded. Of the three renowned algorithms, naive Bayes was shown to achieve the highest improved value when PSO feature selection was used.

3) COMPARATIVE ANALYSIS BASED ON THE ACCURACY MEASUREMENTS OF EACH CLASSIFIER

Our final experiment was to perform a comparative analysis based on the accuracy measurements of each classifier. The aim was to find the true positive and false positive rates and to generate a receiver operating characteristic (ROC) curve. Without proper feature selection, the area under the ROC curve was sparse for all three algorithms. Some key points should be noted here: even after the careful selection of attributes, the area under the ROC curve was below the mark for any algorithm.

The improvement of the ROC curve showed the importance of best feature selection for the datasets. Figure 15, Figure 16, Figure 17, and Figure 18 show every evaluation criteria, which achieved a certain level of improvement when

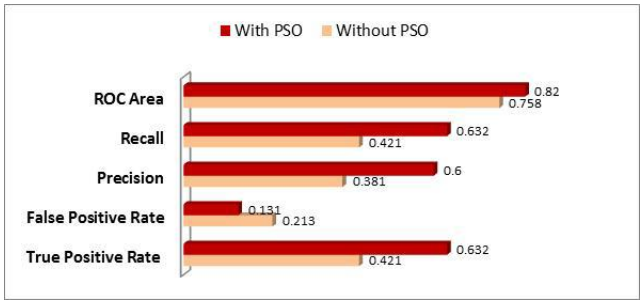

FIGURE 15. Accuracy Measurement for the Naïve Bayes Classifier.

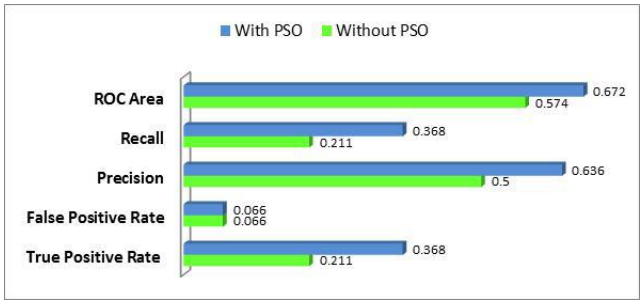

FIGURE 16. Accuracy Measurement for the REPTree Classifier.

FIGURE 17. Accuracy Measurement for the K-Nearest Neighbor Classifier.

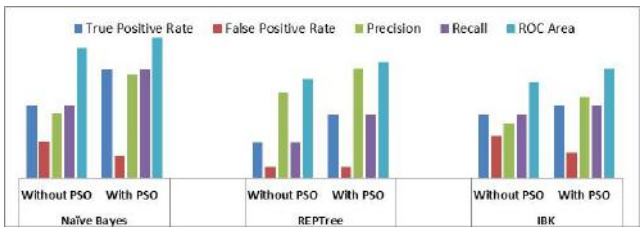

FIGURE 18. Graphical Comparison of Different Classifiers for Different Evaluation Criteria.

appropriate feature selection was confirmed. This result is substantial evidence that a medium-sized dataset with a large number of attributes can easily be misguided by an additional number of features with exceptionally less contribution to classification.

VIII. CONCLUSION AND FUTURE WORK

In this paper, we focused on investigating the effect of integrating the feature selection algorithm with classification algorithms in breast cancer prognosis. We proposed that we can improve most classification algorithms by using feature selection techniques to reduce the number of features. Some features have more importance and influence over the results of the classification algorithms compared to other features. We have presented the results of our experiments on three popular classifying algorithms, namely, naïve Bayes, IBK, and REPTree, with and without the feature selection algorithm, particle swarm optimization (PSO). To conclude, naive Bayes produced better output with and without PSO, whereas the other two techniques improved when used with PSO.

The future direction of this study will include testing newer algorithms with other feature selection techniques. We will experiment on cluster techniques as well as ensemble algorithms.

REFERENCES

[1] R. Siegel, J. Ma, Z. Zou, and A. Jemal, "Cancer statistics, 2014," *CA, Cancer J. Clinicians*, vol. 64, no. 1, pp. 9–29, 2014.

[2] A. Bhardwaj and A. Tiwari, "Breast cancer diagnosis using Genetically optimized neural network model," *Expert Syst. Appl.*, vol. 42, no. 10, pp. 4611–4620, 2015.

[3] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," *CA, A Cancer J. Clinicians*, vol. 61, no. 2, pp. 69–90, 2011.

[4] D. R. Youlden, S. M. Cramb, N. A. Dunn, J. M. Müller, C. M. Pyke, and P. D. Baade, "The descriptive epidemiology of female breast cancer: An international comparison of screening, incidence, survival and mortality," *Cancer Epidemiol.*, vol. 36, no. 3, pp. 237–248, 2012.

[5] N. S. El Saghir *et al.*, "Trends in epidemiology and management of breast cancer in developing Arab countries: A literature and registry analysis," *Int. J. Surg.*, vol. 5, no. 4, pp. 225–233, 2007.

[6] K. Ravichandran and A. S. Al-Zahrani, "Association of reproductive factors with the incidence of breast cancer in Gulf cooperation council countries," *East Mediterr. Health J.*, vol. 15, no. 3, pp. 612–621, 2009.

[7] D. Thompson and D. Easton, "The genetic epidemiology of breast cancer genes," *J. Mammary Gland. Biol. Neoplasia*, vol. 9, no. 3, pp. 221–236, 2004.

[8] F. Bray, P. McCarron, and D. M. Parkin, "The changing global patterns of female breast cancer incidence and mortality," *Breast Cancer Res.*, vol. 6, no. 6, pp. 229–239, 2004.

[9] D. M. Parkin, F. Bray, J. Ferlay, and P. Pisani, "Global cancer statistics, 2002," *CA, Cancer J. Clinicians*, vol. 55, no. 2, pp. 74–108, 2005.

[10] K. McPherson, C. M. Steel, and K. M. Dixon, "Breast cancer—Epidemiology, risk factors, and genetics," *BMJ*, vol. 321, no. 7261, pp. 624–628, 2000.

[11] N. M. Perera and G. P. Gui, "Multi-ethnic differences in breast cancer: Current concepts and future directions," *Int. J. Cancer*, vol. 106, no. 4, pp. 463–467, 2003.

[12] A. A. Ezzat, E. M. Ibrahim, M. A. Raja, S. Al-Sobhi, A. Rostom, and R. K. Stuart, "Locally advanced breast cancer in Saudi Arabia: High frequency of stage III in a young population," *Med. Oncol.*, vol. 16, no. 2, pp. 95–103, Jul. 1999.

[13] E. M. Ibrahim, A. A. Ezzat, M. M. Rahal, M. M. Raja, and D. S. Ajarim, "Adjuvant chemotherapy in 780 patients with early breast cancer: 10-year data from Saudi Arabia," *Med. Oncol.*, vol. 22, no. 4, pp. 343–352, 2005.

[14] N. Elkum *et al.*, "Being 40 or younger is an independent risk factor for relapse in operable breast cancer patients: The Saudi Arabia experience," *BMC Cancer*, vol. 7, p. 222, Dec. 2007.

[15] H. Najjar and A. Easson, "Age at diagnosis of breast cancer in Arab nations," *Int. J. Surg.*, vol. 8, no. 6, pp. 448–452, 2010.

[16] A. Farr, R. Wuerstlein, A. Heiduschka, C. F. Singer, and N. Harbeck, "Modern risk assessment for individualizing treatment concepts in early-stage breast cancer," *Rev. Obstetrics Gynecol.*, vol. 6, nos. 3–4, pp. 165–173, 2013.

[17] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, p. 37, 1996.

[18] M. Seera and C. P. Lim, "A hybrid intelligent system for medical data classification," *Expert Syst. Appl.*, vol. 41, no. 5, pp. 2239–2249, 2014.
[19] 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, 2009.
[20] 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, 2014.
[21] K. Polat, S. Şahan, H. Kodaz, and S. Güneş, "Breast cancer and liver disorders classification using artificial immune recognition system (AIRS) with performance evaluation by fuzzy resource allocation mechanism," *Expert Syst. Appl.*, vol. 32, no. 1, pp. 172–183, 2007.
[22] E. D. Übeyli, "Implementing automated diagnostic systems for breast cancer detection," *Expert Syst. Appl.*, vol. 33, no. 4, pp. 1054–1062, 2007.
[23] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3240–3247, 2009.
[24] J. Thongkam, G. Xu, Y. Zhang, and F. Huang, "Toward breast cancer survivability prediction models through improving training space," *Expert Syst. Appl.*, vol. 36, pp. 12200–12209, Apr. 2009.
[25] N. Sharma and H. Om, "Data mining models for predicting oral cancer survivability," *Netw. Model. Anal. Health Inform. Bioinf.*, vol. 2, no. 4, pp. 285–295, 2013.
[26] 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, 2012.
[27] 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, 2011.
[28] E. Avci, "A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier," *Expert Syst. Appl.*, vol. 36, no. 7, pp. 10618–10626, 2009.
[29] S.-W. 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, 2010.
[30] N. H. Barakat, A. P. Bradley, and M. N. H. Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 4, pp. 1114–1120, Jul. 2010.
[31] 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, Dec. 2015.
[32] C. J. C. Burges, "Dimension reduction: A guided tour," *Found. Trends Mach. Learn.*, vol. 2, no. 4, pp. 275–365, 2010.
[33] T.-H. Cheng, C.-P. Wei, and V. S. Tseng, "Feature selection for medical data mining: Comparisons of expert judgment and automatic approaches," in *Proc. 19th IEEE Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2006, pp. 165–170.
[34] S. N. Ghazavi and T. W. Liao, "Medical data mining by fuzzy modeling with selected features," *Artif. Intell. Med.*, vol. 43, no. 3, pp. 195–206, 2008.
[35] S. M. Vieira, J. M. C. Sousa, and U. Kaymak, "Fuzzy criteria for feature selection," *Fuzzy Sets Syst.*, vol. 189, no. 1, pp. 1–18, 2012.
[36] Z. Pang *et al.*, "A computer-aided diagnosis system for dynamic contrast-enhanced mr images based on level set segmentation and relieff feature selection," *Comput. Math. Methods Med.*, vol. 2015, pp. 1–10, 2015.
[37] T. Liu, L. Hu, C. Ma, Z. Y. Wang, and H.-L. Chen, "A fast approach for detection of erythemato-squamous diseases based on extreme learning machine with maximum relevance minimum redundancy feature selection," *Int. J. Syst. Sci.*, vol. 46, no. 5, pp. 919–931, 2015.
[38] M. Kumar, A. Sharma, and S. Agarwal, "Clinical decision support system for diabetes disease diagnosis using optimized neural network," in *Proc. Students Conf. IEEE Eng. Syst. (SCES)*, vol. 3, Aug. 2014, pp. 254–261.
[39] M. Anbarasi, E. Anupriya, and N. C. S. N. Iyengar, "Enhanced prediction of heart disease with feature subset selection using genetic algorithm," *Int. J. Eng. Sci. Technol.*, vol. 2, no. 10, pp. 5370–5376, 2010.
[40] E. Bron, M. Smits, W. Niessen, and S. Klein, "Feature selection based on the SVM weight vector for classification of dementia," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 5, pp. 1617–1626, Sep. 2015.
[41] T. Santhanam and M. S. Padmavathi, "Application of k-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis," *Proc. Comput. Sci.*, vol. 47, pp. 76–83, Jan. 2015.
[42] Z. Zhang, J. Dong, X. Luo, K.-S. Choi, and X. Wu, "Heartbeat classification using disease-specific feature selection," *Comput. Biol. Med.*, vol. 46, pp. 79–89, Mar. 2014.
[43] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*, 2nd ed. Piscataway, NJ, USA: Wiley, 2011.
[44] C. Címpanu and L. Ferariu, "Survey of data clustering algorithms," Univ. Tehnică, Gheorghe Asachi' din Iasi Tomul LVIII (LXII), SecŃia AUTOMATICÁ si CALCULATOARE Fasc 3, 2012.
[45] T.-P. Hong, K.-Y. Lin, and S.-L. Wang, "Fuzzy data mining for interesting generalized association rules," *Fuzzy Sets Syst.*, vol. 138, no. 2, pp. 255–269, 2003.
[46] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.
[47] B. Liu, Y. Xiao, L. Cao, Z. Hao, and F. Deng, "SVDD-based outlier detection on uncertain data," *Knowl. Inf. Syst.*, vol. 34, no. 3, pp. 597–618, 2013.
[48] D. Prilutsky, B. Rogachev, R. S. Marks, L. Lobel, and M. Last, "Classification of infectious diseases based on chemiluminescent signatures of phagocytes in whole blood," *Artif. Intell. Med.*, vol. 52, no. 3, pp. 153–163, 2011.
[49] B. Samanta *et al.*, "Prediction of periventricular leukomalacia. Part I: Selection of hemodynamic features using logistic regression and decision tree algorithms," *Artif. Intell. Med.*, vol. 46, no. 3, pp. 201–215, 2009.
[50] T. M. Lehmann *et al.*, "Automatic categorization of medical images for content-based retrieval and data mining," *Comput. Med. Imag. Graph*, vol. 29, nos. 2–3, pp. 143–155, 2005.
[51] R. Veloso, F. Portela, M. F. Santos, Á. Silva, F. Rua, A. Abelha, and J. Machado, "A clustering approach for predicting readmissions in intensive medicine," *Proc. Technol.*, vol. 16, pp. 1307–1316, Jan. 2014.
[52] G. Niu, S. Singh, S. W. Holland, and M. Pecht, "Health monitoring of electronic products based on Mahalanobis distance and Weibull decision metrics," *Microelectron. Reliab.*, vol. 51, no. 2, pp. 279–284, 2011.
[53] R. Armañanzas, C. Bielza, K. R. Chaudhuri, P. Martinez-Martin, and P. Larrañaga, "Unveiling relevant non-motor Parkinson's disease severity symptoms using a machine learning approach," *Artif. Intell. Med.*, vol. 58, no. 3, pp. 195–202, 2013.
[54] P. J. García-Laencina, P. H. Abreu, M. H. Abreu, and N. Afonoso, "Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values," *Comput. Biol. Med.*, vol. 59, pp. 125–133, Apr. 2015.
[55] R. Henriques, C. Antunes, and S. C. Madeira, "Generative modeling of repositories of health records for predictive tasks," *Data Mining Knowl. Discovery*, vol. 29, no. 4, pp. 999–1032, 2015.
[56] S. Şahan, K. Polat, H. Kodaz, and S. Güneş, "A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis," *Comput. Biol. Med.*, vol. 37, no. 3, pp. 415–423, 2007.
[57] E. Çomak, K. Polat, S. Güneş, and A. Arslan, "A new medical decision making system: Least square support vector machine (LSSVM) with fuzzy weighting pre-processing," *Expert Syst. Appl.*, vol. 32, no. 2, pp. 409–414, 2007.
[58] H. M. Zolbanin, D. Delen, and A. H. Zadeh, "Predicting overall survivability in comorbidity of cancers: A data mining approach," *Decis. Support Syst.*, vol. 74, pp. 150–161, Jun. 2015.
[59] L. Minelli, M. C. d'Ornellas, and A. T. Winck, "Knowledge representation for lung cancer patients' prognosis," in *Proc. IEEE 16th Int. Conf. E-Health Netw., Appl. Services (Healthcom)*, Natal, Brazil, Oct. 2014, pp. 358–363.
[60] *University of California Irvine Machine Learning Repository*. Accessed: Jan. 1, 2018. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast±Cancer±Wisconsin±(Prognostic)
[61] J. Brownlee. (2017). Machine learning mastery. Machine Learning Mastery, Melbourne, VIC, Australia. [Online]. Available: https://machinelearningmastery.com
[62] C.-L. Huang, H.-C. Liao, and M.-C. Chen, "Prediction model building and feature selection with support vector machines in breast cancer diagnosis," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 578–587, 2008.
[63] B. Amarnath and S. Balamurugan, "Review on feature selection techniques and its impact for effective data classification using UCI machine learning repository dataset," *J. Eng. Sci. Technol.*, vol. 11, no. 11, pp. 1639–1646, 2016.
[64] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," *Proc. Nat. Acad. Sci. USA*, vol. 87, no. 23, pp. 9193–9196, 1990.

[65] F. Provost and R. Kohavi, "Guest editors' introduction: On applied research in machine learning," *Mach. Learn.*, vol. 30, nos. 2–3, pp. 127–132, 1998.
[66] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.
[67] J. R. Quinlan, "Simplifying decision trees," *Int. J. Hum.-Comput. Studies*, vol. 51, no. 2, pp. 497–510, 1999.
[68] S. K. Jayanthi and S. Sasikala, "REPTree classifier for identifying link spam in Web search engines," *ICTACT J. Soft Comput.*, vol. 3, no. 2, pp. 498–505, 2013.
[69] N. Altman and B. MacGibbon, "Consistent bandwidth selection for kernel binary regression," *J. Stat. Planning Inference*, vol. 70, no. 1, pp. 121–137, 1998.
[70] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Adv. Bioinf.*, vol. 2015, pp. 1–13, 2015
[71] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "A new hybrid filter–wrapper feature selection method for clustering based on ranking," *Neurocomputing*, vol. 214, pp. 866–880, Nov. 2016.
[72] R. M. Sharkawy, K. Ibrahim, M. M. A. Salama, and R. Bartnikas, "Particle swarm optimization feature selection for the classification of conducting particles in transformer oil," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 18, no. 6, pp. 1897–1907, Dec. 2011.
[73] B. S. Khehra and A. P. S. Pharwaha, "Comparison of genetic algorithm, particle swarm optimization and biogeography-based optimization for feature selection to classify clusters of microcalcifications," *J. Inst. Eng. (India), B*, vol. 98, no. 2, pp. 189–202, 2017.
[74] M. Xi *et al.*, "Cancer feature selection and classification using a binary quantum-behaved particle swarm optimization and support vector machine," *Comput. Math. Methods Med.*, vol. 2016, pp. 1–9, 2016.
[75] B. Z. Dadaneh, H. Y. Markid, and A. Zakerolhosseini, "Unsupervised probabilistic feature selection using ant colony optimization," *Expert Syst. Appl.*, vol. 53, pp. 27–42, Jul. 2016.
[76] X. Wang, J. Yang, X. Teng, W. Xia, and R. Jensen, "Feature selection based on rough sets and particle swarm optimization," *Pattern Recognit. Lett.*, vol. 28, no. 4, pp. 459–471, 2007.
[77] T. M. Blackwell, "Particle swarms and population diversity," *Soft Comput.*, vol. 9, no. 11, pp. 793–802, 2005.
[78] M. Abdel-Basset, A. E. Fakhry, I. El-henawy, T. Qiu, and A. K. Sangaiah, "Feature and intensity based medical image registration using particle swarm optimization," *J. Med. Syst.*, vol. 41, p. 197, Dec. 2017.
[79] H. Su, "Siting and sizing of distributed generators based on improved simulated annealing particle swarm optimization," in *Environmental Science and Pollution Research*. Berlin, Germany: Springer, 2017, doi: 10.1007/s11356-017-0823-3.
[80] R. Liu, Y. Chen, L. Jiao, and Y. Li, "A particle swarm optimization based simultaneous learning framework for clustering and classification," *Pattern Recognit.*, vol. 47, no. 6, pp. 2143–2152, 2014.
[81] S. Mandal, "A modified particle swarm optimization algorithm based on self-adaptive acceleration constants," *Int. J. Modern Edu. Comput. Sci.*, vol. 9, no. 8, p. 49, 2017.
[82] Z. Hu, Q. Su, and X. Xia, "Multiobjective image color quantization algorithm based on self-adaptive hybrid differential evolution," *Comput. Intell. Neurosci.*, vol. 2016, pp. 1–12, 2016.
[83] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2012.

SAPIAH BINTI SAKRI received the B.Sc. degree in computer science from the National University of Malaysia, Malaysia, in 1985, the M.Sc. degree in information science from the University of Malaya, Malaysia, in 1997, and the Ph.D. degree in information security from the National University of Malaysia, in 2007. Her Ph.D. thesis was on information security management. She was with the Prime Minister's Department, Malaysia, for 30 years. She was the Director of the Information Communication Technology Policy and Strategic Division, Prime Minister's Department. She is currently an Assistant Professor with the Department of Information Systems, Princess Nourah bint Abdulrahman University, Saudi Arabia. Her research interests include information security management, cloud computing security, data mining, machine learning, deep learning, and data science.

NURAINI BINTI ABDUL RASHID received the B.Sc. degree from Mississippi State University, USA, and the M.Sc. and Ph.D. degrees from University Sains Malaysia, Malaysia, all in computer science. Her Ph.D. thesis was on analyzing and managing protein sequence data. She was with the School of Computer Sciences, University Sains Malaysia, for over 25 years. She is currently an Associate Professor with the Department of Computer Sciences, Princess Nourah bint Abdulrahman University, Saudi Arabia. She has published over 70 refereed papers in bioinformatics, string matching, and information retrieval. Her research interests include parallel algorithms, information retrieval methods, and biological big data mining.

ZUHAIRA MUHAMMAD ZAIN received the bachelor's degree in information science and the master's degree in computer science from the National University of Malaysia in 2000 and 2007, respectively, and the Ph.D. degree in software engineering from Universiti Putra Malaysia in 2013. She was a Programmer with Fujitsu Systems Global Solutions Sdn Bhd from 2000 to 2007. She is currently an Assistant Professor with the College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Her research interests include software engineering, software quality, software evaluation, and data analytics.