electronics

MDPI

Article

A Hybrid Supervised Machine Learning Classifier System for Breast Cancer Prognosis Using Feature Selection and Data Imbalance Handling Approaches

Yogendra Singh Solanki [1], Prasun Chakrabarti [2], Michal Jasinski [3,*], Zbigniew Leonowicz [3], Vadim Bolshev [4,*], Alexander Vinogradov [4], Elzbieta Jasinska [5], Radomir Gono [6] and Mohammad Nami [7]

[1] Lincoln University College, No. 2, Jalan Stadium, SS 7/15, Kelana Jaya, 47301 Petaling Jaya, Malaysia; yogendra.phd@lincoln.edu.my

[2] Department of Computer Science Engineering, Techno India NJR Institute of Technology, Udaipur, Rajasthan 313003, India; drprasun.cse@gmail.com

[3] Department of Electrical Engineering Fundamentals, Faculty of Electrical Engineering, Wroclaw University of Science and Technology, 50-370 Wroclaw, Poland; zbigniew.leonowicz@pwr.edu.pl

[4] Laboratory of Power Supply and Heat Supply, Federal Scientific Agroengineering Center VIM, 109428 Moscow, Russia; schkolamolen@gmail.com

[5] Faculty of Law, Administration and Economics, University of Wroclaw, 50-145 Wroclaw, Poland; elzbieta.jasinska@uwr.edu.pl

[6] Department of Electrical Power Engineering, Faculty of Electrical Engineering and Computer Science, VSB—Technical University of Ostrava, 708 00 Ostrava, Czech Republic; radomir.gono@vsb.cz

[7] Department Department of Neuroscience, Shiraz University of Medical Sciences, Shiraz 71348-14336, Iran; leanmtneurosci2@gmail.com

* Correspondence: michal.jasinski@pwr.edu.pl (M.J.); vadimbolshev@gmail.com (V.B.); Tel.: +48-713-202-022 (M.J.); +7-499-174-85-95 (V.B.)

Abstract: Nowadays, breast cancer is the most frequent cancer among women. Early detection is a critical issue that can be effectively achieved by machine learning (ML) techniques. Thus in this article, the methods to improve the accuracy of ML classification models for the prognosis of breast cancer are investigated. Wrapper-based feature selection approach along with nature-inspired algorithms such as Particle Swarm Optimization, Genetic Search, and Greedy Stepwise has been used to identify the important features. On these selected features popular machine learning classifiers Support Vector Machine, J48 (C4.5 Decision Tree Algorithm), Multilayer-Perceptron (a feed-forward ANN) were used in the system. The methodology of the proposed system is structured into five stages which include (1) Data Pre-processing; (2) Data imbalance handling; (3) Feature Selection; (4) Machine Learning Classifiers; (5) classifier's performance evaluation. The dataset under this research experimentation is referred from the UCI Machine Learning Repository, named Breast Cancer Wisconsin (Diagnostic) Data Set. This article indicated that the J48 decision tree classifier is the appropriate machine learning-based classifier for optimum breast cancer prognosis. Support Vector Machine with Particle Swarm Optimization algorithm for feature selection achieves the accuracy of 98.24%, MCC = 0.961, Sensitivity = 99.11%, Specificity = 96.54%, and Kappa statistics of 0.9606. It is also observed that the J48 Decision Tree classifier with the Genetic Search algorithm for feature selection achieves the accuracy of 98.83%, MCC = 0.974, Sensitivity = 98.95%, Specificity = 98.58%, and Kappa statistics of 0.9735. Furthermore, Multilayer Perceptron ANN classifier with Genetic Search algorithm for feature selection achieves the accuracy of 98.59%, MCC = 0.968, Sensitivity = 98.6%, Specificity = 98.57%, and Kappa statistics of 0.9682.

Keywords: breast cancer prognosis; supervised machine learning classifier; data selection; imbalance handling

check for updates

Citation: Solanki, Y.S.; Chakrabarti, P.; Jasinski, M.; Leonowicz, Z.; Bolshev, V.; Vinogradov, A.; Jasinska, E.; Gono, R.; Nami, M. A Hybrid Supervised Machine Learning Classifier System for Breast Cancer Prognosis Using Feature Selection and Data Imbalance Handling Approaches. *Electronics* **2021**, *10*, 699. https://doi.org/10.3390/electronics10060699

Academic Editors: Christian Morbidoni, Francesco Di Nardo and Alessandro Cucchiarelli

Received: 10 February 2021
Accepted: 12 March 2021
Published: 16 March 2021

1. Introduction

Breast cancers are the most frequent cancers among women, according to World Health Organization. It concerns 2.1 million women each year, and it also causes the greatest number of cancer-related deaths of women [1,2]. In India Breast cancer is the most common form of cancer. In metro cities like Mumbai, Delhi, Bangalore breast cancer accounts for 25% to 32% of female cancers. This condition becomes more serious because nowadays it became more noticeable in the younger age groups. Around 50% of all cases are in the age group of range between 25 and 50 [3]. The numbers are shocking and constantly rising [4,5]. According to the Indian Council for Medical Research in 2016, the total number of new cancer cases was about 14.5×10^5 and this figure is likely to increase to 17.3×10^5 in 2020. As the number of breast cancer cases in India increases, cancer fear levels increase too. If it's not able to prevent breast cancer, it can increase the survival rates by being informed and choosing the right treatment at the right time. To improve breast cancer outcomes and survival, early detection is critical which can be effectively achieved by machine learning (ML) and data mining techniques [6,7]. The ML algorithms such as classification techniques can be utilized to develop a model to diagnose breast cancer either as malignant or benign [8,9]. Various data mining techniques such as class balancing, re-sampling, etc. can be used to handle the dataset and improve the classification accuracy. Once the data imbalance has been handled then using the same by applying feature selection algorithms, we can obtain the most important features which play important role in the accuracy of the classification model as well as reduce the computation time. Many such approaches have been proposed and we used nature-inspired algorithms.

This computation is done on breast cancer datasets on available repository datasets from the University of California, Irvine. We have implemented different classification methods to classify the data to detect the malignant and benign groups from the given dataset and applied various imbalance data handling techniques and feature selection algorithms to improve the performance of the classifiers. To classify breast cancer cells as malignant or benign by ML classifiers, many researchers have worked around. Saoud et al. [10] have examined six different ML techniques for breast cancer diagnosis and found that Bayes network and support vector machine (SVM) gave an accuracy of 97.2818% on the Wisconsin breast cancer dataset. Saoud et al [11] proposed an approach for breast cancer detection using supervised and unsupervised machine learning algorithms and showed that supervised algorithms are more efficient. Domingo et al. [12] analyzed the various decision trees for classifying breast cancer stages. They observed that the fuzzy decision tree had better performance than the J48 tree. Sahu et al. [13] found out that SVM was more efficient in comparison to other techniques, and studied the parameters such as accuracy, specificity, and sensitivity. Al-Shargabi et al. [14] have obtained the best result for breast cancer classification with K-Nearest Neighbors and Random Forest with an accuracy of 100%, the second rank for the original Multi-Layer Perceptron with an accuracy of 97.19%. Zhang et al. [15] have addressed the diagnosis of breast cancer and class imbalance problem using the K-Boosted C5.0 algorithm based on under-sampling. Devi et al. [16] performed a comparative analysis among various ML algorithms evaluated based on the basis accuracy and ROC curve of each classifier. Fotouhi et al. [17] have examined oversampling and under-sampling on various cancer datasets and found that balancing techniques had improved the classification of cancer datasets.

Many researchers have worked on feature selection approaches to further improve the accuracy of ML Classifiers. al Haq et al. [18] suggested the hybrid framework using ML classifiers with Relief, Lasso, and mRMR feature selection algorithms for heart disease. Ahmed Abdullah Farid etal. [19] have proposed an early diagnosis system for breast cancer using the CHFS feature selection algorithm and SVM achieved 98.25% accuracy. Bibhuprasad Sahu etal. [20] have proposed the cancer classification approach based on SVM optimized with particle swarm optimization and reverse firefly algorithm. Sahu etal. [21] have proposed the predictive model of cancer diagnosis using multivariate statistical and machine learning techniques for better accuracy. Kewat et al. [22] have evaluated wrapper-based feature selection techniques particle swarm optimization, genetic search, and greedy stepwise. Tabrizchi et al. [23] have proposed breast cancer diagnosis using the multi-verse optimizer-based gradient boosting decision tree. They have combined Gradient Boosting Decision Tree (GBDT) and multi-verse optimizer (MVO) to propose a robust classifier for optimal classification.

Based on the literature review, it was observed that in the majority of the cases various classifiers and feature selection approaches had been used to improve accuracy. In this paper, we propose a framework with a hybrid approach, which extends [24] where the accuracy improvement by handling the data imbalance using re-sampling and SMOTE (Synthetic Minority Over-sampling Technique) technique has been suggested. The same technique has been used along with nature-inspired feature selection approaches to improve the accuracy of the classification models. In this work, the accuracy evaluation of ML classifiers has been done based on parameters such as Kappa statistics and MCC (Matthews Correlation Coefficient) which had rarely been taken into consideration in other literature.

Lahoura et al. [25] have used techniques based on artificial neural networks (ANN) to verify, the possibility to apply it to disease diagnosis. The extreme learning machine is an example of ANN. It has a huge potential to solve various classification issues. The proposed paper approach is based on amalgamates three research domains. Firstly, an extreme learning machine was used to diagnose breast cancer. Then, the gain ratio feature selection method was used to eliminate insignificant features. Finally, the cloud computing-based system for remote diagnosis was proposed. The obtained results indicated that accuracy achieved is around 0.987, recall is 0.913, precision is 0.905, and F1-score is 0.813.

Yu et al. [26] compared RMAF and RELU and other activation functions on deeper models. The RMAF was selected as the most appreciated. Experiments were based on training and classification on multi-layer perceptron MLP by benchmarking data. The applied dataset concerns Wisconsin breast cancer, MNIST, Iris, and Car evaluation. The results of the RMAF investigation indicated that the performance of 98.74%, 99.67%, 98.81%, and 99.42%. Then it was compared to Sigmoid, Tanh, and ReLU. Then, the experiment concerned the convolution neural network using MNIST, CIFAR-10, and CIFAR-100 data. The indicated performance accuracy was 99.73%, 98.77%, and 79.82% in comparison to Tanh, ReLU, and Swish.

Ferreira et al. [27] distinguished five types of cancer. The investigation concerned RNA-Seq datasets: thyroid, skin, stomach, breast, and lung. Then the performance comparison was based on three autoencoders applied as a deep neural network weight initialization technique.

This work is segregated as follows: Section 2 presents details about the materials and proposed methodology applied in this research. Section 2.1 describes the details about the dataset under consideration. Section 2.2 entails in detail the methodology of the proposed system, which includes feature selection algorithms, ML classifiers, evaluation parameters, etc. Section 3 presents and discusses the results obtained and the comparison of the results with different approaches applied as proposed in the methodology. Section 4 discusses the conclusion of the conducted experiment and suggests the methodology to obtain improved accuracy by implementing the hybrid approach.

The research has a huge social impact as it will facilitate medical treatment through the prognosis of breast cancer at its early stages. The related accuracy rate can be noted that will further be used as the threshold for future treatment of breast cancer. The research will be extremely helpful for academicians, researchers, oncologists and the results will lead to novel techniques for the proper prognosis of breast cancer.

2. Materials and Methods

The following subsections briefly discuss the research materials and methodology used for this paper.

2.1. Dataset and Tools

The dataset under research experimentation is referred from UCI Machine Learning Repository, named Breast Cancer Wisconsin (Diagnostic) Data Set [28]. The dataset consists of features computed from digital images of fine needle aspirate of a breast mass. These features represent the characteristics of cell nuclei present in the image. The dataset attribute description is represented in Table 1. It consists of 32 different features on cell images of 569 participants out of which 63% cases are benign and 37% belong to malignant. It shows that this dataset consists of data imbalance due to which the accuracy of classifiers affects a lot. The same has already been handled in the previous research paper [24].

Table 1. Description of dataset attributes.

S.No.	Feature Name
1	ID Number
2	Diagnosis (M = Malignant, B = Benign)
	(3–32) Ten real-valued features for each cell nucleus
A	Radius (mean of distances from the center to points on the perimeter)
B	Texture (standard deviation of gray-scale values)
C	Perimeter
D	Area
E	Smoothness (local variation in radius lengths)
F	Compactness (perimeterˆ2 / area − 1.0)
G	Concavity (severity of concave portions of the contour)
H	Concave points (number of concave portions of the contour)
I	Symmetry
J	Fractal dimension (“coastline approximation” − 1)

All the experiments were performed on a personal computer with the following specifications: 64 bit Intel Core2Duo 2.93 GHz processor, 6 GB RAM. Weka 3.8.4 as a classification tool.

2.2. Methodology for the Proposed System

The proposed system has been designed to classify malignant cells from benign ones. In this research, we worked on methods to enhance the accuracy of machine learning classification models for the prognosis of Breast cancer. The performance of the classifiers has been tested on all attributes and selected features separately to obtain and compare the achieved accuracy. Wrapper-based feature selection approach along with nature-inspired algorithms such as Particle Swarm Optimization (PSO), Genetic Search, and Greedy Stepwise have been used to identify the important features. On these selected features popular machine learning classifiers Support Vector Machine (SVM), J48 (C4.5 Decision Tree Algorithm), Multilayer-Perceptron (a feed-forward ANN) were used in the system. The methodology of the proposed system is structured into five stages which include: (1) Data Pre-processing; (2) Data imbalance handling; (3) Feature Selection; (4) Machine Learning Classifiers; (5) classifier's performance evaluation. Figure 1 shows the proposed framework.

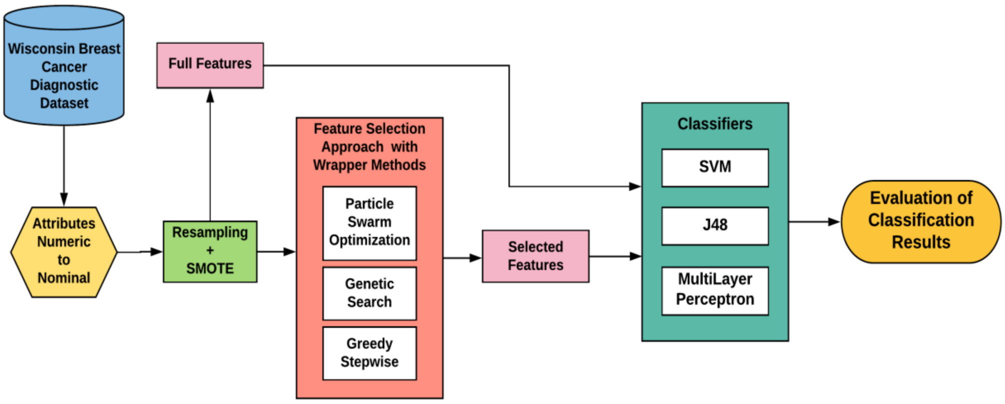

Wisconsin Breast Cancer Diagnostic Dataset
Full Features
Attributes Numeric to Nominal
Resampling + SMOTE
Feature Selection Approach with Wrapper Methods
Particle Swarm Optimization
Genetic Search
Greedy Stepwise
Selected Features
Classifiers
SVM
J48
MultiLayer Perceptron
Evaluation of Classification Results

Figure 1. Proposed framework for breast cancer classification.

2.2.1. Data Pre-Processing

Data pre-processing is required for efficient data representation. This stage includes the removal of missing values, conversion of data from numeric to nominal, which is the discretization of features. By performing discretization, it is easy for the decision tree to create branches that are easy to understand rather than branches based on numbers. The missing values feature row is removed from the dataset.

2.2.2. Data Imbalance Handling

Although handling data imbalance may be considered as a part of data pre-processing, we mention it separately here. What is data imbalance? It is the condition in the dataset where the number of instances per class is not equally distributed, which leads to misleading classification accuracy. To improve the performance of classifiers and handling imbalance conditions, we can use any of the following approaches under-sampling, over-sampling, generation of synthetic samples. In the proposed framework we have used oversampling with SMOTE method.

2.2.3. Feature Selection Algorithms

Feature Selection Algorithm is one more important step in the machine learning classification process [29], as most of the time, there are many features in the dataset which are irrelevant or have the least correlation with the output classes for example serial or ID number in any dataset. Such features affect the performance of the machine learning classifiers. Feature selection improves classification accuracy and reduces model execution time [30,31].

Particle Swarm Optimization

Particle Swarm Optimization (PSO) [32,33] is a metaheuristic algorithm based on the concept inspired by swarm behavior such as bird flocking in nature. It was proposed by Kennedy and Eberhart in the year 1995. The indicated algorithm emulates the interaction between members to share information. PSO was used in different areas e.g., optimization and combination with other existing algorithms. The PSO method concerns a search of the optimal solution by agents. They are referred to as particles, that trajectories are adjusted by both the stochastic and deterministic components. Each particle is influenced by the best-obtained position and the best position of the group. Finally, it tends to move randomly.

Genetic Search

Genetic Algorithm (GA) is an example of a search-based optimization technique. It is based on the principles of Genetics and Natural Selection. It was inspired by Charles Darwin's evolution theory. John Holland and his students and colleagues at the University of Michigan were the developers of this approach. David E. Goldberg worked on various optimization problems. In GAs, there is a pool with possible solutions for the given problem. These solutions then undergo recombination and mutation, to produce new children, and the process is repeated. Each solution is assigned a fitness value based on its objective function value and based on that the fitter ones are chosen to yield more "fitter" solutions. In this way, we keep "evolving" better solutions over generations until we reach a stopping criterion [34–36].

Greedy Stepwise

Greedy Stepwise is a forward stepwise elimination where we start with finding the variable that maximizes the accuracy and then keeps on increasing the number of variables as long as the accuracy of the model increases (i.e., greedy-search optimization) [37]. There are, however, issues with this approach. The first one is that it does not check all the combinations. It evaluates them one by one. However, a variable that may not work well standalone may lead to higher accuracies with the interaction of another variable. The second issue is the selection of the prediction algorithm to use while selecting the variables. Conventionally, computationally cheap linear regression is preferred.

2.2.4. Machine Learning Classifiers

Once the data were pre-processed where all the anomalies had been handled, they are now passed on to the machine learning classifiers to train an ML model which could classify cancer in a breast cell as malignant or benign. In this section, brief information about the machine learning classifiers considered for the research is discussed [38].

Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm. It can be applied to classification or regression tasks. However, it is preferred for classification problems. In the SVM algorithm, each data item is plotted as a point in x-dimensional space (where x is equal to the number of features) with the value of each feature being the value of the particular coordinate. As a solution to separate the two classes of the data points, many possible hyperplanes may be applied. Here the objective is to find a plane that has the maximum distance between data points of each class. By a maximization of a margin distance, it is provided with some reinforcement so that future data points can be classified with more confidence. The loss function that helps maximize the margin is hinge loss.

J48

J48 represents the open-source Java implementation of the C4.5 algorithm. It is the algorithm applied to generate a decision tree, that was developed by Ross Quinlan. It is an extension of Quinlan's earlier ID3 algorithm. In this case, the decision trees are generated using C4.5 for classification, and for this reason, C4.5 is often referred to as the statistical classifier. It selects one attribute from a set of training instances and then selects an initial subset of the training instances. Now the attribute and the subset of instances are used to build a decision tree. The rest of the training instances (those not in the subset used for construction) is used to test the accuracy of the constructed tree. It will be iterated until a tree is built. C4.5 uses the information gain ratio to select the attribute which best differentiates the instances.

Multilayer Perceptron

Multilayer Perceptron (MLP) is representative of a deep artificial neural network. It is composed of more than one perceptron. These are composed using an input layer to receive the signal. Then the output layer makes a decision or prediction about the input. Finally, in between those two, an arbitrary number of hidden layers is noticeable. They are used as the true computational engine of the MLP. Multilayer perceptrons are used to train on a set of input-output pairs and learn to model the correlation (or dependencies) between those both inputs and outputs. The training activities involve adjusting the parameters, or the weights and biases, of the model to minimize error level. Back-propagation is used to make those weight and bias adjustments relative to the error. The error itself can be measured in a variety of ways, including by root mean squared error (RMSE).

2.2.5. Performance Evaluation Metrics

The true positive represents the outcome that the model correctly predicts the positive class. The True negative represents the outcome that the model correctly predicts the negative class. The false-positive represents the outcome that the model incorrectly predicts the positive class. The false-negative represents the outcome that the model incorrectly predicts the negative class.

Classification Accuracy

The classification accuracy of an ML classifier is the solution to measure how often the algorithm classifies a data point correctly. The accuracy informs about the number of correctly predicted data points out of all the data points, which is evaluated as follows:

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{\text{TP} + \text{TN} \ + \text{FP} + \text{FN}} \tag{1}$$

Analyzing only the accuracy is not sufficient to deal with a class-imbalanced data set, where their significant disparity is noticed between the number of positive and negative labels.

Sensitivity

The test sensitivity is named the true positive rate (TPR). It concerns the proportion of samples that are genuinely positive that give a positive result using the test in question. It also concerns type II errors; false negatives are the failures to reject a false null hypothesis.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{FN} + \text{TP}} \tag{2}$$

Specificity

The test specificity is named the true negative rate (TNR). It concerns the proportion of samples that test negative using the test in question that are genuinely negative. Additionally, it is referred to as type I errors, false positives are the rejection of a true null hypothesis. It is evaluated as follows:

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{3}$$

Matthew's Correlation Coefficient

Matthew's Correlation Coefficient (MCC) ranges from -1 to 1. The -1 means a completely inaccurate binary classifier. The value 1 means a completely correct binary classifier. The application of the MCC enables one to gauge how well their classification model is performed. Unlike the F1 score (F-score, known also as F1-score), represents the measure of a dataset model's accuracy. It is applied to assure the evaluation of binary classification systems. Those systems classify examples into "positive" as well as "negative". The F1-score enables the combination of the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall. The F1-score is applied to assure the evaluation of information retrieval systems such as search engines, and also for many kinds of machine learning models). MCC is represented as a single-value metric, it summarizes a confusion matrix. A confusion matrix (or error matrix), has four entries: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). MCC concerns the true class and the predicted class as two (binary) variables. Then it computes their correlation coefficient the higher level of the correlation between true and predicted values means that it is a better prediction. This is the phi-coefficient (φ), which is renamed as the Matthews Correlation Coefficient (MCC), and evaluated as follows:

$$\mathrm{MCC} = \frac{((\mathrm{TN} \times \mathrm{TP}) - (\mathrm{FP} \times \mathrm{FN}))}{\sqrt{(\mathrm{TN}+\mathrm{FN})(\mathrm{FP}+\mathrm{TP})(\mathrm{TN}+\mathrm{FP})(\mathrm{FN}+\mathrm{TP})}} \quad (4)$$

Kappa Statistics

Cohen's kappa statistic measures the inter-rater reliability, which means it is the agreement between two raters who each classify the N items into C mutually exclusive classes. Its value range is 0–1. Reference [39] the formula for evaluation of Cohen's Kappa coefficient is as follows:

$$\mathrm{K} = \frac{\mathrm{P0} - \mathrm{Pe}}{1 - \mathrm{Pe}} \quad (5)$$

where

P0: Probability of agreement.
Pe: Probability of random agreement

3. Results

This section of the paper involves the discussion on ML classification models and results obtained with different approaches. First of all, we applied the ML classifiers on the pre-processed data in which data imbalance was handled using the re-sampling and SMOTE approach. The result of this experiment with detailed accuracy is shown in Table 2 and for comparison, the percentage accuracy has been plotted in Figure 2. In the second step, we used the nature-inspired feature selection algorithms (like Particle Swarm Optimization, Genetic Search, and Greedy Stepwise) along with the Wrapper methods like (Naïve Bayes, KNN, J48 decision tree, and Random forest) on both the dataset one with preprocessing and one without preprocessing. The number of features selected by these approaches is shown in Figure 3. It can be observed from the table as well as from the figure that on average the number of features selected by various feature selection algorithms with wrapper evaluation functions is lesser when applied on data after pre-processing rather than data without preprocessing. After applying feature selection, we applied machine learning classifiers like SVM, Decision tree J48, and Multilayer perceptron on both the processed and unprocessed datasets. The accuracy comparison of three classifiers with the PSO approach is shown in Table 3, similarly, accuracy comparison with the genetic search algorithm and Greedy stepwise is shown in Tables 4 and 5, respectively. The accuracy of different classifiers with different feature selection approaches is compared in Figures 4 and 5.

Table 2. Performance of ml classifiers on data after pre-processing.

Classifiers	SVM	J48	MultiLayer Perceptron
Correctly Classified Instances (Total 852)	835	568	840
Accuracy (%)	98.00	66.67	98.59
MCC	0.955	N.A	0.969
Sensitivity (%)	98.761	66.667	99.467
Specificity (%)	96.51	N.A	96.89
AUC	0.979	0.494	0.997
PRC Area	0.971	0.553	0.997
Kappa statistic	0.9552	0.000	0.9685
Mean absolute error	0.020	0.444	0.0165
Root mean squared error	0.1413	0.474	0.1184
Relative absolute error (%)	4.49	99.97	3.71
Root relative squared error (%)	29.96	100.00	25.11

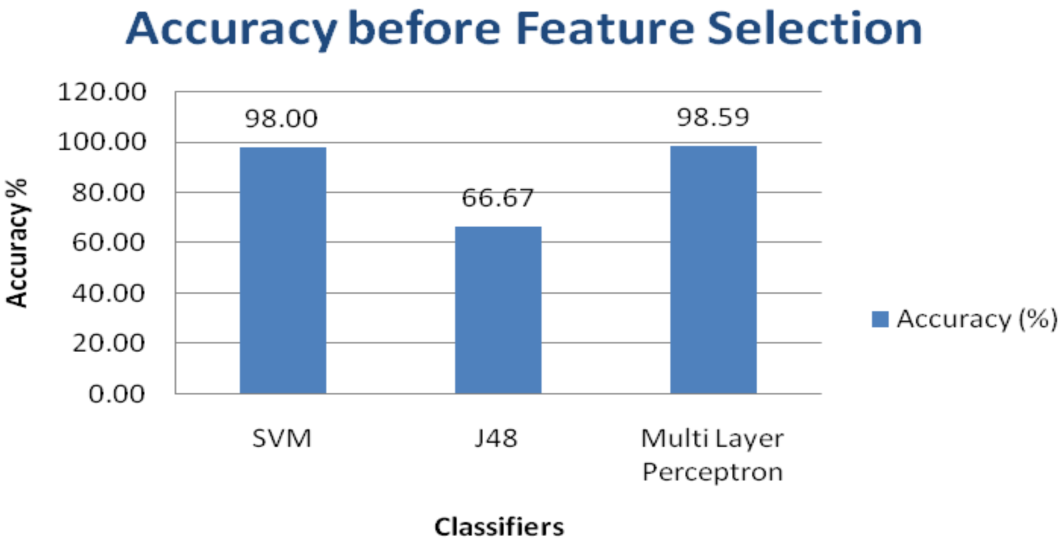


Figure 2. Comparative accuracy of multilayer (ML) classifiers on data after pre-processing.

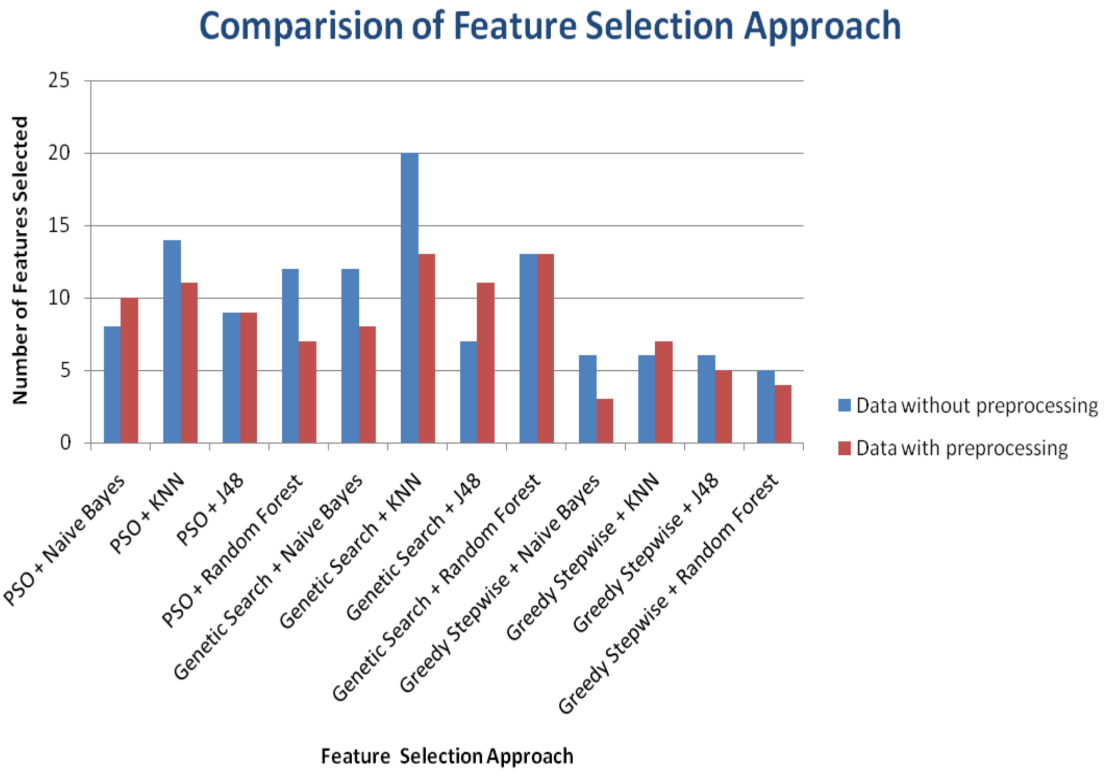


Figure 3. Comparison of the number of features selected.

Table 3. Accuracy comparison of classifiers on data with particle swarm optimization (PSO) feature selection.

Classifiers	Wrapper Based Feature Selection Approach	Number of FeaturesSelected		Accuracy(%)		Time to Build the Model (Seconds)	
		WoDP	WDP	WoDP	WDP	WoDP	WDP
SVM	PSO + Naive Bayes	8	10	97.0123	97.6526	0.16 s	0.53 s
	PSO + KNN	14	11	96.6608	98.2394	0.08 s	0.03 s
	PSO + J48	9	9	96.6608	95.7746	0.02 s	0.04 s
	PSO + RandomForest	12	7	95.9578	92.2535	0.01 s	0.02 s
J48	PSO + Naive Bayes	8	10	94.9033	98.0047	0.03 s	0.03 s
	PSO + KNN	14	11	94.2004	98.1221	0.01 s	0.02 s
	PSO + J48	9	9	96.6608	98.3568	0.01 s	0.02 s
	PSO + RandomForest	12	7	94.3761	98.2394	0.01 s	0.03 s
Multi-Layer Perceptron	PSO + Naive Bayes	8	10	96.4851	98.0047	0.65 s	0.95 s
	PSO + KNN	14	11	96.3093	97.5352	1.00 s	0.98 s
	PSO + J48	9	9	96.4851	97.0657	0.52 s	0.74 s
	PSO + Random Forest	12	7	97.0123	97.4178	0.79 s	0.60 s

WoDP—without data processing, WDP—with data processing, s—Seconds.

Table 4. Accuracy comparison of classifiers on data with genetic search feature selection.

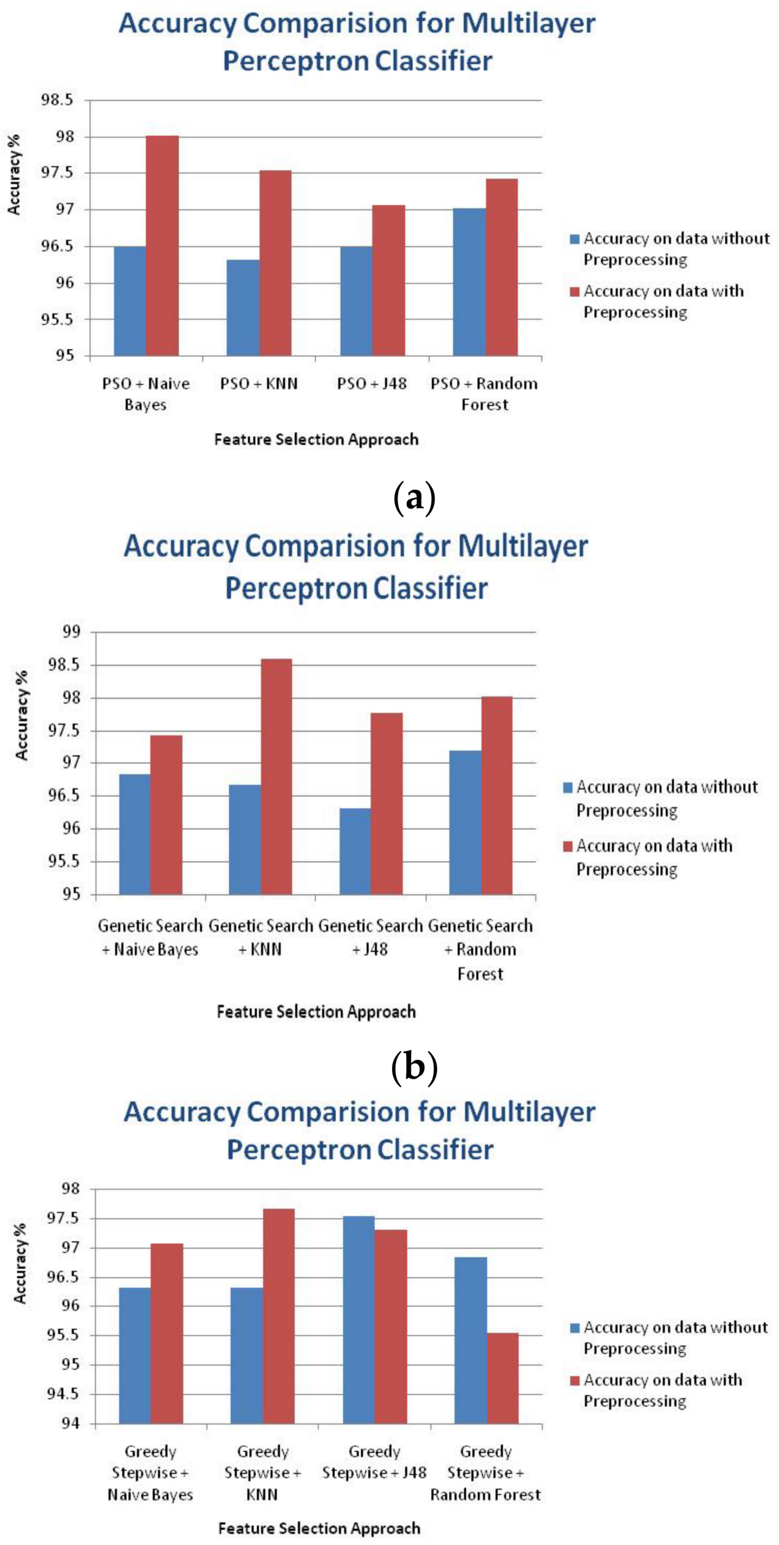
Classifiers	Wrapper Based Feature Selection Approach	Number of Features Selected		Accuracy (%)		Time to Build the Model (Seconds)	
		WoDP	WDP	WoDP	WDP	WoDP	WDP
SVM	PSO + Naive Bayes	12	8	97.0123	97.1831	0.02 s	0.02 s
	PSO + KNN	20	13	97.188	97.5352	0.03 s	0.02 s
	PSO + J48	7	11	95.9578	96.3615	0.02 s	0.02 s
	PSO + Random Forest	13	13	97.5395	97.7700	0.01 s	0.02 s
J48	PSO + Naive Bayes	12	8	94.9033	97.7700	0.01 s	0.01 s
	PSO + KNN	20	13	94.3761	97.5352	0.02 s	0.02 s
	PSO + J48	7	11	96.1336	98.8263	0.01 s	0.01 s
	PSO + Random Forest	13	13	94.9033	98.8263	0.01 s	0.02 s
Multi-Layer Perceptron	PSO + Naive Bayes	12	8	96.8366	97.4178	0.79 s	0.72 s
	PSO + KNN	20	13	96.6608	98.5915	1.63 s	1.23 s
	PSO + J48	7	11	96.3093	97.7700	0.39 s	1.05 s
	PSO + Random Forest	13	13	97.188	98.0047	0.83 s	1.23 s

WoDP—without data processing, WDP—with data processing, s—Seconds.

Table 5. Accuracy comparison of classifiers on data with greedy stepwise feature selection.

Classifiers	Wrapper Based Feature Selection Approach	Number of Features Selected		Accuracy (%)		Time to Build the Model (Seconds)	
		WoDP	WDP	WoDP	WDP	WoDP	WDP
SVM	PSO + Naive Bayes	6	3	96.1336	96.7136	0.01 s	0.01 s
	PSO + KNN	6	7	96.6608	97.4178	0.01 s	0.03 s
	PSO + J48	6	5	95.4306	97.0657	0.02 s	0.03 s
	PSO + Random Forest	5	4	96.4851	94.6009	0.02 s	0.02 s
J48	PSO + Naive Bayes	6	3	94.9033	96.9484	0.01 s	0.05 s
	PSO + KNN	6	7	95.4306	98.3568	0.01 s	0.01 s
	PSO + J48	6	5	97.0123	98.5915	0.01 s	0.01 s
	PSO + Random Forest	5	4	95.9578	97.0657	0.01 s	0.01 s
Multi-Layer Perceptron	PSO + Naive Bayes	6	3	96.3093	97.0657	0.37 s	0.27 s
	PSO + KNN	6	7	96.3093	97.6526	0.39 s	0.58 s
	PSO + J48	6	5	97.5395	97.3005	0.37 s	0.42 s
	PSO + Random Forest	5	4	96.8366	95.5399	0.32 s	0.40 s

WoDP—without data processing, WDP—with data processing, s—Seconds.

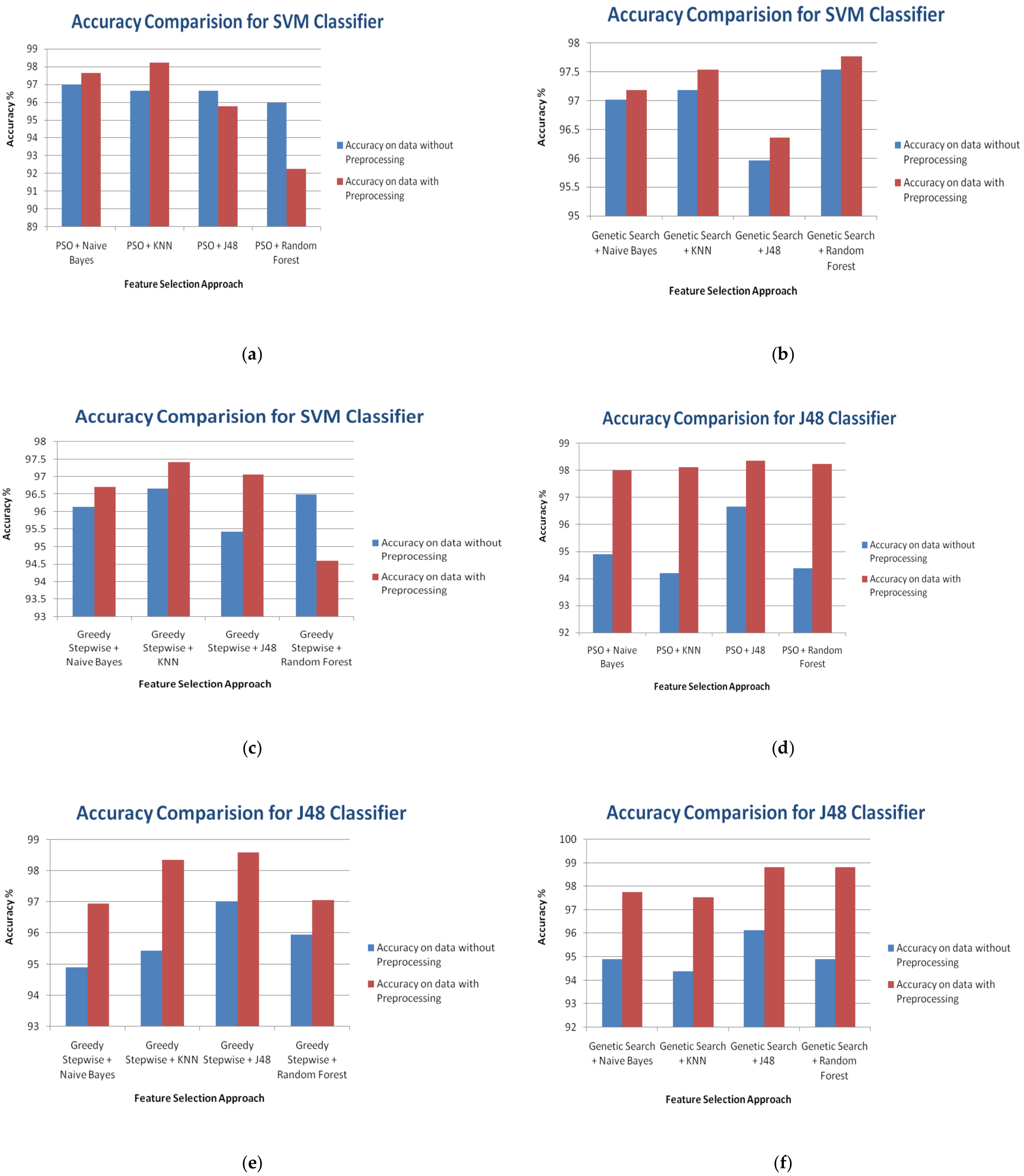


(**a**)

(**b**)

(**c**)

Figure 4. Accuracy comparison of Multilayer Perceptron Classifier for (**a**) PSO features selection algorithm; (**b**) Genetic Search features selection algorithm; (**c**) Greedy Stepwise features selection algorithm.



(**a**) (**b**)

(**c**) (**d**)

(**e**) (**f**)

Figure 5. Accuracy comparison of support vector machine (SVM) for (**a**) PSO features selection algorithm; (**b**) Genetic Search features selection algorithm; (**c**) Greedy Stepwise features selection algorithm; (**d**) Accuracy comparison of J48 for PSO features selection algorithm; (**e**) Genetic Search features selection algorithm; (**f**) Greedy Stepwise features selection algorithm.

4. Discussion

Following are the observations from the accuracy comparison from Figures 4 and 5. In most cases, the accuracy of ML classifiers with the features from data with pre-processing is better than the features selected from data without preprocessing. This indicates that feature selection on data after applying re-sampling and SMOTE improves the accuracy of the classifier. For SVM classifier three out of all feature selection algorithm, PSO with KNN evaluator gives the maximum accuracy of 98.24%. Likewise, for the J48 Decision tree, the maximum accuracy of 98.83% is achieved by a Genetic search with a J48 evaluator. For the Multilayer perceptron classifier, the highest accuracy of 98.59% is obtained by using a Genetic search algorithm with KNN. Accuracy details of all the above-mentioned classifiers with other performance evaluation parameters such as MCC, Sensitivity, specificity AUC, Kappa statistics, etc. are shown in Table 6. It is observed that out of all J48 decision tree classifiers with Genetic search feature selection algorithm outperforms all other classifiers not only in terms of accuracy but also in terms of Mathew's Coefficient and Cohen's Kappa statistics along with sensitivity and specificity.

Table 6. Comparative accuracy evaluation of classifiers with various feature selection approach. On data after preprocessing with resampling + smote.

FeatureSelection Approach	Classifiers	Correctly Classified Instances (Total 852)	Accuracy (%)	MCC	Sensitivity (%)	Specificity (%)	AUC	PRC Area	Kappa Statistic	Meanabsolute Error	Root Meansquared Error	Relativeabsolute Error (%)	Root Relativesquared Error (%)
PSO + Naive Bayes	SVM	832	97.7	0.95	98.8	95.5	0.98	0.97	0.95	0.02	0.15	5.3	32.5
	J48	835	98.0	0.96	98.4	97.2	0.97	0.96	0.96	0.03	0.14	5.7	29.9
	MP	835	98.0	0.96	98.9	96.2	1.00	1.00	0.96	0.03	0.13	6.2	27.1
PSO + KNN	SVM	837	98.2	0.96	99.1	96.5	0.98	0.98	0.96	0.02	0.13	4.0	28.1
	J48	836	98.1	0.96	98.4	97.5	0.98	0.97	0.96	0.02	0.14	5.1	28.9
	MP	831	97.5	0.95	98.2	96.1	0.99	0.99	0.94	0.03	0.14	5.9	29.6
PSO + J48	SVM	816	95.8	0.91	98.4	91.1	0.96	0.94	0.91	0.04	0.21	9.5	43.6
	J48	838	98.4	0.96	98.3	98.6	0.98	0.98	0.96	0.03	0.13	5.7	26.8
	MP	827	97.1	0.93	97.5	96.1	0.99	0.99	0.93	0.04	0.16	8.4	34.6
PSO + Random Forest	SVM	786	92.3	0.83	96.1	85.4	0.92	0.90	0.83	0.08	0.28	17.4	59.0
	J48	837	98.2	0.96	98.4	97.9	0.98	0.98	0.96	0.02	0.13	5.3	28.1
	MP	830	97.4	0.94	98.6	95.2	1.00	1.00	0.94	0.04	0.14	8.3	29.5
Genetic Search + Naive Bayes	SVM	828	97.2	0.94	97.4	96.8	0.97	0.96	0.94	0.03	0.17	6.3	35.6
	J48	833	97.8	0.95	98.2	96.8	0.98	0.98	0.95	0.02	0.14	5.6	30.5
	MP	830	97.4	0.94	98.8	94.9	0.99	0.99	0.94	0.03	0.14	7.8	30.0
Genetic Search + KNN	SVM	831	97.5	0.95	98.8	95.2	0.98	0.97	0.94	0.02	0.16	5.5	33.3
	J48	831	97.5	0.94	97.7	97.1	0.97	0.97	0.94	0.03	0.15	6.1	32.8
	MP	840	98.6	0.97	98.6	98.6	0.99	0.99	0.97	0.02	0.12	4.4	25.0
Genetic Search + J48	SVM	821	96.4	0.92	97.4	94.4	0.96	0.95	0.92	0.04	0.19	8.2	40.5
	J48	842	98.8	0.97	98.9	98.6	0.98	0.98	0.97	0.02	0.11	3.8	22.8
	MP	833	97.8	0.95	97.7	97.8	1.00	1.00	0.95	0.03	0.14	6.1	29.7
Genetic Search + Random Forest	SVM	833	97.8	0.95	98.9	95.5	0.98	0.97	0.95	0.02	0.15	5.0	31.7
	J48	842	98.8	0.97	99.1	98.2	0.98	0.98	0.97	0.02	0.11	4.0	22.9
	MP	835	98.0	0.96	99.1	95.9	1.00	1.00	0.96	0.02	0.12	4.8	25.6
Greedy Stepwise + Naive Bayes	SVM	824	96.7	0.93	96.9	96.4	0.96	0.95	0.93	0.03	0.18	7.4	38.5
	J48	826	96.9	0.93	97.2	96.4	0.98	0.98	0.93	0.04	0.17	8.3	36.0
	MP	827	97.1	0.93	97.9	95.4	1.00	1.00	0.93	0.04	0.14	8.9	30.7
Greedy Stepwise + KNN	SVM	830	97.4	0.94	97.4	97.5	0.97	0.96	0.94	0.03	0.16	5.8	34.1
	J48	838	98.4	0.96	98.8	97.5	0.98	0.98	0.96	0.02	0.13	4.7	27.0
	MP	832	97.7	0.95	98.4	96.2	0.99	0.99	0.95	0.03	0.14	7.6	30.1
Greedy Stepwise + J48	SVM	827	97.1	0.93	98.1	95.1	0.97	0.96	0.93	0.03	0.17	6.6	36.3
	J48	840	98.6	0.97	99.1	97.6	0.99	0.98	0.97	0.02	0.12	3.6	24.8
	MP	829	97.3	0.94	98.6	94.8	1.00	1.00	0.94	0.04	0.14	8.6	30.4
Greedy Stepwise + Random Forest	SVM	806	94.6	0.88	96.4	91.0	0.94	0.92	0.88	0.05	0.23	12.1	49.3
	J48	827	97.1	0.93	97.9	95.4	0.97	0.97	0.93	0.03	0.17	7.7	35.5
	MP	814	95.5	0.90	98.0	91.0	0.99	0.99	0.90	0.07	0.18	14.8	38.7

Based on the Feature selection approach (PSO + Naive Bayes), the classifiers J48 and Multilayer Perceptron deliver the best accuracy (98.0%). In feature selection based on (PSO + KNN), the accuracy rate of the SVM classifier is best (98.2%). Based on the feature selection approach (PSO + J48), J48 delivers the best accuracy of 98.4%. Based on the feature selection approach (PSO + Random Forest), J48 delivers the best accuracy of 98.2%. With a feature selection approach (Genetic Search + Naive Bayes), J48 delivers the best accuracy of 97.8%. With (Genetic Search + KNN) approach Multilayer Perceptron gives the best accuracy of 98.6%. Based on the feature selection approach (Genetic Search + Random Forest) and (Genetic Search + J48), J48 delivers the best accuracy of 98.8%. Based on the feature selection approach (Greedy Stepwise + Naive Bayes), Multilayer perceptron delivers the best accuracy of 97.1%. In the feature selection approach (Greedy Stepwise + KNN), (Greedy Stepwise + J48), and (Greedy Stepwise + Random Forest), the J48 classifier delivers the best accuracy of >97%. The Kappa statistics is maximum (0.973) in J48 classifier based on Genetic Search feature selection algorithm. The error rate is also minimum in J48 thereby yielding the best accuracy rate of 98.83%. The sensitivity and specificity are maximum in the case of SVM (99.11%) and J48 (98.58%), respectively. The work can be further assessed based on a 95% confidence interval.

5. Conclusions

The paper points out a Hybrid Supervised Machine Learning Classifier System for breast cancer prognosis using feature selection and data imbalance approaches. The performance of the classifiers has been tested on all attributes and selected features separately to obtain and compare the achieved accuracy. Wrapper-based feature selection approach along with nature-inspired algorithms such as Particle Swarm Optimization, Genetic Search, and Greedy Stepwise has been used to identify important features. On these selected features popular machine learning classifiers such as Support Vector Machine, J48 (C4.5 Decision Tree Algorithm), Multilayer-Perceptron (a feed-forward ANN) were used in the system. The methodology of the proposed system is structured into five stages which include (1) data pre-processing; (2) data imbalance handling; (3) feature selection; (4) machine learning classifiers; (5) classifier's performance evaluation. Based on the experimental results, it is evident that the Support Vector Machine with the Particle Swarm Optimization algorithm for feature selection achieves an accuracy of 98.24%, MCC of 0.961, a sensitivity of 99.11%, a specificity of 96.54%, and Kappa statistics of 0.9606. It is also observed that the J48 Decision Tree classifier with the Genetic Search algorithm for feature selection achieves an accuracy of 98.83%, MCC of 0.974, a sensitivity of 98.95%, a specificity of 98.58%, and Kappa statistics of 0.9735. Furthermore, Multilayer Perceptron ANN classifier with Genetic Search algorithm for feature selection achieves the accuracy of 98.59%, MCC of 0.968, a sensitivity of 98.6%, a specificity of 98.57%, and Kappa statistics of 0.9682. Given the above, it is relevant that the J48 decision tree classifier is the most appropriate machine learning-based classifier for optimum breast cancer prognosis. This work will facilitate medical treatment towards breast cancer prognosis in the light of machine learning. The future scope of work includes the prognosis of breast cancer using thermal images and IoT-based sensors.

Author Contributions: Conceptualization, Y.S.S., methodology, P.C.; software, Y.S.S. and P.C.; validation, Z.L., A.V. and R.G.; formal analysis, M.J., E.J. and V.B.; investigation, Y.S.S.; resources, P.C.; data curation, M.J. and M.N.; writing—original draft preparation, Y.S.S. and P.C.; writing—review and editing, M.J. and V.B.; visualization, Y.S.S. and E.J.; supervision, P.C., Z.L., A.V. and R.G.; project administration, M.J.; funding acquisition, Z.L. and E.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wu, J.; Mamidi, T.K.K.; Zhang, L.; Hicks, C. Unraveling the Genomic-Epigenomic Interaction Landscape in Triple Negative and Non-Triple Negative Breast Cancer. *Cancers* **2020**, *12*, 1559. [CrossRef] [PubMed]
2. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2019. *CA. Cancer J. Clin.* **2019**, *69*, 7–34. [CrossRef] [PubMed]
3. Gupta, G.; Dang, R.; Gupta, S. Clinical presentations of carcinoma breast in rural population of North India: A prospective observational study. *Int. Surg. J.* **2019**, *6*, 1622–1635. [CrossRef]
4. Kalarivayil, R.; Desai, P.N. Emerging technologies and innovation policies in India: How disparities in cancer research might be furthering health inequities? *J. Asian Public Policy* **2020**, *13*, 192–207. [CrossRef]
5. Raina, V.; Deo, S.; Shukla, N.; Mohanti, B.; Gogia, A. Triple-negative breast cancer: An institutional analysis. *Indian J. Cancer* **2014**, *51*, 163–178. [CrossRef]
6. Roy, S.; Kumar, R.; Mittal, V.; Gupta, D. Classification models for Invasive Ductal Carcinoma Progression, based on gene expression data-trained supervised machine learning. *Sci. Rep.* **2020**, *10*, 4113–4126. [CrossRef]
7. Saba, T. Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges. *J. Infect. Public Health* **2020**, *13*, 1274–1289. [CrossRef] [PubMed]
8. Chakravarthy, S.R.S.; Rajaguru, H. Detection and classification of microcalcification from digital mammograms with firefly algorithm, extreme learning machine and non-linear regression models: A comparison. *Int. J. Imaging Syst. Technol.* **2020**, *30*, 126–146. [CrossRef]
9. Eedi, H.; Kolla, M. Machine Learning aproaches for healthcare data analysis. *J. Crit. Rev.* **2020**, *7*, 312–326. [CrossRef]
10. Saoud, H.; Ghadi, A.; Ghailani, M.; Boudhir, A.A. Application of data mining classification algorithms for breast cancer diagnosis. *ACM Int. Conf. Proc. Ser.* **2018**, *20*, 34–46. [CrossRef]
11. Saoud, H.; Ghadi, A.; Ghailani, M. Proposed approach for breast cancer diagnosis using machine learning. *ACM Int. Conf. Proc. Ser.* **2019**, *21*, 1–5. [CrossRef]
12. Domingo, M.J.; Gerardo, B.D.; Medina, R.P. Fuzzy decision tree for breast cancer prediction. *ACM Int. Conf. Proc. Ser.* **2019**, *12*, 316–328. [CrossRef]
13. Sahu, B.; Panigrahi, A. Efficient Role of Machine Learning Classifiers in the Prediction and Detection of Breast Cancer. *SSRN Electron. J.* **2020**, *10*, 1–9. [CrossRef]
14. Al-Shargabi, B.; Al-Shami, F. An experimental study for breast cancer prediction algorithms. *ACM Int. Conf. Proc. Ser.* **2019**, *21*, 3–8. [CrossRef]
15. Zhang, J.; Chen, L.; Abid, F. Prediction of Breast Cancer from Imbalance Respect Using Cluster-Based Undersampling Method. *J. Healthc. Eng.* **2019**, *2019*. [CrossRef]
16. Prabadevi, B.; Deepa, K.L.B.N.; Vinod, V. Analysis of Machine Learning Algorithms on Cancer Dataset. In Proceedings of the 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 24–25 February 2020; pp. 1–10. [CrossRef]
17. Fotouhi, S.; Asadi, S.; Kattan, M.W. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *J. Biomed. Inform.* **2019**, *90*, 103089. [CrossRef] [PubMed]
18. Haq, A.U.; Li, J.P.; Memon, M.H.; Nazir, S.; Sun, R.; Garciá-Magarinõ, I. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mob. Inf. Syst.* **2018**, *2018*. [CrossRef]
19. Farid, A.A.; Selim, G.I.; Khater, H.A. A Composite Hybrid Feature Selection Learning-Based Optimization of Genetic Algorithm for Breast Cancer Detection. *Preprints* **2020**, *25*, 1–21. [CrossRef]
20. Sahu, B.; Panigrahi, A.; Mohanty, S.; Sobhan, S. A hybrid Cancer Classification Based on SVM Optimized by PSO and Reverse Firefly Algorithm. *Int. J. Control Autom.* **2020**, *13*, 506–517.
21. Sahu, B.; Mohanty, S.N.; Rout, S.K. EAI Endorsed Transactions on Scalable Information System s A H ybrid Approach for Breast Cancer Classification and Diagnosis. *EAI Endorsed Trans. Scalable Inf. Syst.* **2019**, *21*, 1–8. [CrossRef]
22. Kewat, A.; Srivastava, P.N.; Kumhar, D. Performance Evaluation of Wrapper-Based Feature Selection Techniques for Medical Datasets. *Algorithms Intell. Syst.* **2020**, *32*, 619–633. [CrossRef]
23. Tabrizchi, H.; Tabrizchi, M. Breast cancer diagnosis using a multi-verse optimizer-based gradient boosting decision tree. *SN Appl. Sci.* **2020**, *2*, 1–19. [CrossRef]
24. Solanki, Y.S.; Chakrabarti, P. Analysis of Breast Cancer Prognosis Using Supervised Machine Learning Classifiers. *Int. J. Adv. Sci. Technol.* **2020**, *29*, 10262–10269.
25. Yu, Y.; Adu, K.; Tashi, N.; Anokye, P.; Wang, X.; Ayidzoe, M.A. RMAF: Relu-Memristor-Like Activation Function for Deep Learning. *IEEE Access* **2020**, *8*, 72727–72741. [CrossRef]
26. Lahoura, V.; Singh, H.; Aggarwal, A.; Sharma, B.; Mohammed, M.; Damaševičius, R.; Kadry, S.; Cengiz, K. Cloud Computing-Based Framework for Breast Cancer Diagnosis Using Extreme Learning Machine. *Diagnostics* **2021**, *11*, 241. [CrossRef] [PubMed]
27. Ferreira, M.F.; Camacho, R.; Teixeira, L.F. Using autoencoders as a weight initialization method on deep neural networks for disease detection. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 1–18. [CrossRef]
28. William, H.; Wolberg, W.; Street, N.; Olvi, L. Mangasarian. In *UCI Machine Learning Repository*; School of Information and Computer Science, University of California: Irvine, CA, USA, 1995; Available online: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic) (accessed on 10 February 2021).

29. Salappa, A.; Doumpos, M.; Zopounidis, C. Feature selection algorithms in classification problems: An experimental evaluation. *Optim. Methods Softw.* **2007**, *22*, 199–212. [CrossRef]
30. Darzi, M.; AsgharLiaei, A.; Hosseini, M.; Asghari, H. Feature selection for breast cancer diagnosis: A case-based wrapper approach. *World Acad. Sci. Eng. Technol.* **2011**, *53*, 1142–1145.
31. Kwak, N.; Choi, C.-H. Input feature selection for classification problems. *IEEE Trans. Neural Netw.* **2002**, *13*, 143–159. [CrossRef] [PubMed]
32. Ozcan, E.; Mohan, C.K. Analysis of a Simple Particle Swarm Optimization System. *Intell. Eng. Syst. Artif. Neural Netw.* **1998**, *8*, 253–258.
33. Poli, R.; Kennedy, J.; Blackwell, T. Particle swarm optimization. *Swarm Intell.* **2007**, *1*, 33–57. [CrossRef]
34. Lanzi, P. Fast feature selection with genetic algorithms: A filter approach. In Proceedings of the 1997 IEEE International Conference on Evolutionary Computation (ICEC '97), Indianapolis, IN, USA, 13–16 April 1997; pp. 537–540.
35. Punch, W.F.; Goodman, E.D.; Enbody, R.J. *Further Research on Feature Selection and Classification Using Genetic Algorithms*; Springer: Berlin/Heidelberg, Germany, 1993.
36. Vafaie, H.; Imam, I.F. Feature Selection Methods: Genetic Algorithms vs. Greedy-like Search. In Proceedings of the 3rd International Conference on Fuzzy and Intelligent Control Systems, Louisville, KY, USA, 18–21 December 1994.
37. Dag, H.; Sayin, K.E.; Yenidogan, I.; Albayrak, S.; Acar, C. Comparison of feature selection algorithms for medical data. In Proceedings of the 2012 International Symposium on Innovations in Intelligent Systems and Applications, Trabzon, Turkey, 2–4 July 2012. [CrossRef]
38. Kotsiantis, S.B. Supervised Machine Learning: A Review of Classification Techniques. *Informatica* **2007**, *31*, 249–268.
39. Viera, A.J.; Viera, J.M.G.A.J.; Garrett, J.M. Understandings inter-observer agreement: The kappa statistic. *Fam. Med.* **2005**, *37*, 360–363. [PubMed]