Neural Processing Letters manuscript No.
(will be inserted by the editor)

BCD-WERT: A Novel Approach for Breast Cancer Detection Using Whale Optimization based Efficient Features and Extremely Randomized Tree Algorithm

Shafaq Abbas · Aqsa Akbar · Abdul Rehman Javed · Iqra Batool · Zunera Jalil · G Thippa Reddy

Received: date / Accepted: date

Abstract Breast cancer is one of the leading causes of death in women. It often results in subpar living conditions of a patient as they have to go through with expensive and painful treatments to fight this cancer. Almost half a million women annually do not survive this fight and die from this disease. Early detection of breast cancer is a major factor to detect and cure this disease at the early stages and save precious lives. For this purpose, the medical world has turned their interests towards machine learning algorithms for the detection and prevention of this cancer. Machine learning algorithms have proven to perform accurately in the prediction of breast cancer by building models using the previously available data. In this paper, we propose a novel approach named *BCD-WERT* that utilizes Extra Tree (EXT) and Whale Optimization Algorithm (WOA) for efficient feature selection and Support Vector Machine (SVM), Random Forest (RF), Kernal Support Vector Machine (KSVM), Decision Tree (DT), Logistic Regression (LR), Stochastic Gradient Descent (SDG), Gaussian Naive Bayes (GNB) and k-Nearest Neighbour (KNN) for cancer detection. We performed detailed experiments on

Shafaq Abbas
Department of Computer Science, Faculty of Computing & AI, Air University, Islamabad
E-mail: shafaqabbas54@gmail.com

Aqsa Akbar
Department of Computer Science, Faculty of Computing & AI, Air University, Islamabad
E-mail: 200037@students.au.edu.pk

Abdul Rehman Javed
Department of Cyber Security, Air University, E-9, Islamabad, Pakistan
E-mail: abdulrehman.cs@au.edu.pk

Iqra Batool
Department of Computer Science, Faculty of Computing & AI, Air University, Islamabad
E-mail: 200118@students.au.edu.pk

Zunera Jalil
Department of Cyber Security, Air University, E-9, Islamabad, Pakistan
E-mail: zunera.jalil@mail.au.edu.pk

G Thippa Reddy
School of Information Technology and Engineering, Vellore Institute of Technology, India
E-mail: thippareddy.g@vit.ac.in

state-of-the-art dataset and results shows that the BCD-WERT approach achieves the highest accuracy rate of 99.30% using optimized Extremely Randomized Tree (EXT) algorithm.

Keywords Breast cancer · Support vector machine · Whale optimization algorithm · Feature selection · Machine Learning

1 Introduction

Breast cancer is a life-threatening disease that affects the lives of millions of women. It is also the second leading cause of deaths occurring due to cancer among females [23, 18]. One in eight women is affected by this disease. According to Nalini et al., breast cancer annually affects one million people and results in over 400,000 deaths globally [23]. It is a benign or malignant tumor, caused by uncontrolled growth and division of cells inside the breast [11]. The non-cancerous tumors are called benign which are not life-threatening and can be treated with medicine. The cancerous tumors are known as malignant and if left untreated, they can cause the death of the patient. These tumors appear as a lump in the breast and can be diagnosed using X-rays. For early detection of these tumors, the patient must immediately consult with their healthcare provider if such a lump appears in their breast [23, 7].

Family history, age, genes, dietary habits, and lifestyle are some of the leading factors of this disease. According to Kamel et al., the number of breast cancer cases was higher in high-income countries so far but nowadays, women in low-income countries are also getting sick from this disease [18]. Treatments costs are very high for breast cancer and lately, there are reports of severe side effects after long-term use of medications used to treat this cancer. So, early detection may prove vital in saving the life of a patient. Many of the deaths occur due to the detection of cancer at a stage where it becomes impossible to cure this disease [18, 7].

Recently, medical practitioners have developed a greater interest in the prediction and detection of breast cancer using algorithms of machine learning. Machine learning is a subset of Artificial intelligence that provides the ability to learn and improve from past experiences without being explicitly programmed [21]. Machine learning algorithms such as Decision trees, SVM, Random forest, Naïve Bayes, and K-nearest neighbors are used in prediction and provide remarkable results in medical and other fields [19, 2]. The accuracy rate, computational complexity, and long run-time are some of the limitations of these machine learning algorithms. These can be improved by careful evaluation and a combination of several optimization techniques available [18]. The performance of the classification algorithms can increase tremendously by using feature selection algorithms such as grey wolf optimization, particle swarm optimization, and whale optimization algorithm [32]. Feature selection is used to eliminate any unnecessary data from the dataset and helps improve the accuracy rate of the prediction algorithm. [18, 7].

In this paper, a new hybrid classification technique is introduced. The proposed approach for breast cancer prediction consists of a whale optimization algorithm, which is used to reduce the high dimensionality of the dataset by performing feature selection. Selected features are given to classification algorithms to improve their classification accuracy. The whale optimization algorithm is used for feature selection and Then, the selected features are fed to nine different classification

algorithms for breast cancer prediction. These algorithms include extra tree,SVM, KNN, KSVM, Gaussian Naïve Bayes, stochastic gradient descent, decision tree, Logistic regression, and random forest.

The rest of the paper is structured as follows: section 2 provides a brief literature review of the several feature selection and data mining techniques used for prediction. In section 3, our proposed hybrid approach is discussed in detail. A brief introduction to the whale optimization algorithm and classifiers, SVM, KNN, Kernel SVM, Gaussian Naïve Bayes, random forest, SGD, logistic regression, decision tree, extra tree is discussed. The experimental results of the proposed approach are provided in section 4. Discussion of results and study is done in Section 5. Section 6 presents the conclusion and future work of this study respectively.

2 Literature Review

Breast cancer takes the lives of almost half a million women each year[23]. It should be cured as early as possible. there is a lot of medical treatment for this fatal disease but some machine learning algorithms are also helping in the medical field to predict the diseases and to predict this fatal disease there are some prediction methods proposed. these methods used different datasets and predicted breast cancer.

2.1 Feature Selection based Methods

Al-Zoubi et al proposed a hybrid technique based on SVM and Whale Optimization Algorithm. The model performed spam detection and provides insight into which features play a deciding role in the detection of spam. The proposed model was used on several datasets in a different context (i.e Arabic, English, Spanish, and Korean). It achieved 99% accuracy on Arabic language datasets, 91% on the Spanish language, 96% on the English language, and 95% on the Korean Language dataset [6]. Kamel et al, utilized a combination of Grey Wolf Optimization (GWO) and SVM for breast cancer prediction. They conducted experiments in MATLAB platform and the UCI dataset was used for training and testing. It was found that SVM-GWO produced a 100% accuracy rate whereas 99.29% accuracy rate was achieved without feature selection. The results were compared with other classification algorithms and it was found that in terms of sensitivity, accuracy, and specificity, SVM-GWO outperformed all of them and produced 100 percent results. As compared to work done in literature, the proposed method increased diagnosis accuracy by 27.68 % [18].

Seyed et al, proposed a classification technique based on a whale optimization algorithm and support vector machine for prediction of breast cancer In this model, the KNN classifier was used to extract a subset of features with the best fitness value. WOA was used for the selection of optimal features among the obtained subset. Finally, these best features were fed to SVM classifier for classification of instances. MATLAB platform was used to conduct these experiments. This model achieved an accuracy of 98.77% [29]. Muhammad Nadeemi et al, proposed a meta-heuristic algorithm named “Feature Selection based on Whale Optimization Algorithm (FSWOA)”. FSWOA was used to reduce the dimensionality of medical

data. The accuracy of the proposed FSWOA observed on several medical datasets was 87.10% for Hepatitis, 97.86 percent for Breast Cancer, 78.57% for Pima Indians Diabetes, and 77.05% for Starlog Disease [25].

Sakri et al, compared the results of several algorithms on predicting breast cancer recurrence. This study used PSO as feature selection in combination with renowned classifiers fast decision tree learner, KNN, and Naive Bayes Results of the experiment showed that without(Particle Swam Optimization) PSO feature selection, the highest accuracy achieved was 76.3%. With PSO feature selection, Naïve Bayes provided the highest accuracy that was 81.3% [27]. Sharifi A. et al., Used EM to analyze the data and after normalizing, the Neural network multilayer perceptron structure with WOA was used to predict the breast cancer. The accuracy achieved after performing preprocessing and reducing dimensions of the dataset was 99% and it comes out to be a good machine learning method in comparison to other techniques used [30]. Hiba asri et al, discussed the performance of four classifiers C4.57, SVM5, NB6, and k-NN8 to predict breast cancer. The main purpose of the study was to evaluate the performance of algorithms based on accuracy, sensitivity, and specificity. SVM had the highest accuracy of 97% according to the experiment and all the experiments were conducted using the WEKA tool. The experiments were performed on the Wisconsin dataset[8]. Sivakami et al, compared classifications techniques for breast cancer prediction in Weka, NB, SMO, IBL, Instance-based learning, and DTSVM outperformed all of them and the results obtained have higher prediction accuracy. This work proposed a hybrid methodology to predict the fatal disease and to alarm the consequences of the disease. two methods were used to predict the status of the disease, option extraction, and Information treatment and the other one is the Decision Tree Support Vector machine [31].

Ashutosh et al, used K mean algorithm to predict breast cancer at early stages, the dataset used for prediction was the Breast Cancer Wisconsin dataset. Centroids and distance measures were used to compute the results. Positive prediction accuracy of 92% was achieved. Different correlation techniques were used and, Manhattan and Euclidean proved to be more effective than Pearson correlation, and also according to the results it can be seen that K mean algorithm can be used to classify the BCW dataset [14].

2.2 Data Mining based Methods

Nalini et al, conducted a comparative analysis of data mining methods for breast cancer prediction based on execution time and classification accuracy performance measures. This study used Naïve Bayes and J48 data mining algorithms using open source WEKA tool. The performance of Naïve Bayes and J48 was compared by classification accuracy and execution time. The results show that Naïve Bayes had an accuracy of 64% whereas J48 had an accuracy of 60%. This study concluded that Naïve Bayes is a better classification algorithm for breast cancer prediction because it has a higher accuracy rate and less execution time as compared to J48 [23]. Khourdifi et al., conducted the comparison experiment between K-NN, SVM, Random Forest (RF), and Naïve Bayes to find out the highest accuracy. The result showed that SVM produced the highest accuracy 97.9% [21].

Alghunaim et al, conducted a comprehensive comparative study exploring how well machine learning algorithms perform in breast cancer prediction using big data in terms of performance, effectiveness, and efficiency. Both Spark and WEKA platforms were used to study the difference between their performances as they provide scalable and non-scalable environments respectively. The performance of SVM, RF, and DTs were compared on DNA Methylation (DM), Gene Expression (GE), and their combined datasets. The results show that SVM outperformed decision tree and random forest using all three datasets with 99.68%, 98.73%, and 97.33% accuracy respectively. This study found that GE dataset is the best choice among the three datasets to accurately predict breast cancer[7]. Al-Zoubi et al., [5]conducted experiments to compare the performance of Naïve Bayes C4.5, SVM, and KNN. The results show that SVM is the best classifier and gave the highest accuracy which was 96.99% [5]. Table 1 shows the comparison of the previous studies and the solutions proposed.

Table 1: Literature Review Summary

Authors	**Problem Solved**	**Limitations**
Alghunaim et al. [7]	Cancer Prediction by using SVM ,RF, and DT.	Random forest does not gives precise prediction values.
Kamel et al. [18]	Cancer prediction by using combination of SVM and GWO	Grey Wolf Optimization algorithm has bad local searching ability.
Gouda I. Salama et al. [28]	MLP and J48 classifiers with WBC dataset.	MLP is fully connected that's why it Includes too many parameters.
Gouda I. Salama [28] et al.	SMO and MLP for breast cancer prediction by using the WDBC dataset.	SMO's performance with SVM is not good when the data set has more noise.
Ashutosh kumar dobey et al. [14]	K-means algorithm was used to predict Breast cancer by using the BCW dataset.	Using K-mean with foggy clusters can cause a difference in final clusters
K.Sivakami et al. [31]	Comparison of three classifications for Breast Cancer Prediction.	Degraded performance with increased number of features.
Hiba Asri et al. [8]	SVM, NB, k-NN, and C4.5 on WBC dataset for Breast Cancer	C4. 5 classifier with SVM faces overfitting.
Md.Rezaul Karim et al. [20]	Naïve Bayes and J48 data mining algorithmsfor Breast cancer prediction.	Feature selection can lead to cost effectiveness but J48 and NB in this study did not use the selected feature.

3 BCD-WERT

In this study, a classification based method named as BCD-WERT is proposed for breast cancer prediction. This is a novel approach that uses Whale Optimization based efficient features and well-known classification algorithms. BCD-WERT consists of four main phases. In phase I, the dataset is acquired for experiments. The Wisconsin breast cancer diagnostic dataset from Kaggle online dataset library is used in this study. In phase II, preprocessing is done on the acquired dataset, data is cleaned and brought into a form that is suitable for prediction. In phase III, Whale Optimization Algorithm (WOA) is used for feature selection. In phase IV, selected features is used for the training of the classification models and different classification methods are used for the classification of testing data. For this research study, the whale optimization algorithm and all other classification models are implemented using Python. Figure 1presents the working of the proposed approach.

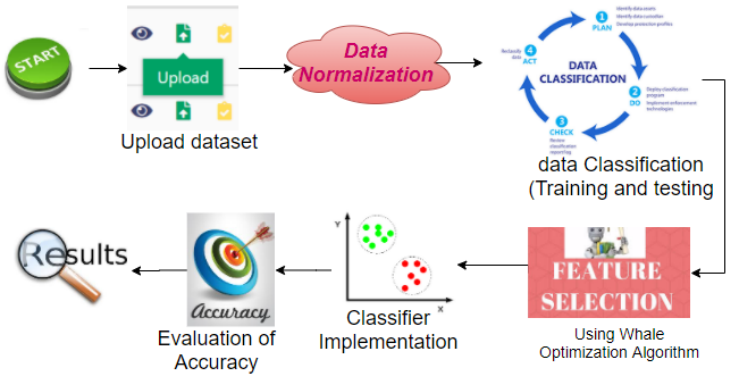


Fig. 1: Overview of Proposed BCD-WERT Approach

3.1 Data Acquisition

The dataset used in this study is acquired from the online Kaggle dataset library [1]. the dataset used is Wisconsin Breast Cancer Diagnostic dataset (WBCD). It contains 569 instances and has 30 attributes. There are two classes an instance can belong to, they are Malignant and Benign. The cancerous class is represented by 1 and non-cancerous is represented by 0 in the dataset. Cancer can be Malignant when cell growth is uncontrollable. The spreading and growth of the cancerous cells can become life-threatening. Malignant tumors rapidly grow and can quickly become part of the other body parts. The WBCD is used various times to predict cancer. The motivation behind using this dataset was to improve the accuracy and

Algorithm 1 Proposed Solution Algorithm

Input: $Reading \leftarrow DatasetReadings$
Output: Benign, Malign
Evaluation Measures: Accuracy, F-Score, Recall, Precision
$i \leftarrow [Reading]$ {Current Instance}
$T \leftarrow []$ {Total Instances}
$P \leftarrow []$ {Predicted Confidence}
$C \leftarrow []$ {Targeted Confidence}
$L \leftarrow [Benign, Malign]$ { Target Class Labels}
Find Best Feature by using WOA
Initialize the population Yj(1,2,3,...,n)
Initialize x,P and z
Calculate the best feature fitness for each search value
Y= the best Search value
Function WOA(population,x,P,z,Maxiter)
i=1
while i $\leq$ Maxiter **do**
  for Each Search Value **do**
    if P$\geq$1 **then**
      Update the position of current search value
    else if P$\geq$1 **then**
      Choose random search value Y
      Update the position of current Value.
    end if
  end for
  Update x,P,z
  Update Yj if Y got better Solution
  I=i+1
end while
Return Yj
End Function
Function Classifiers(Feature_Selected)
Extract Best feature form dataset
Do Computation on selected feature
Processing
while loop until margin constraints violating points **do**
  K splits {sp1,....spk} sp(i) is a random split
  Return split sp* score(sp*)
end while
return yi
Compute Accuracy and confusion matrix

test different classification methods on this dataset to predict malignant breast cancer.

3.2 Data Preprocessing

After the data is acquired, preprocessing[26] is performed on the selected dataset. Data normalization is used to clean the data, by removing irrelevant and incomplete records. This is done to make sure data is consistent and the dataset contains no missing values. After data normalization, randomly splitting of the dataset is done into two subsets i.e. training data and testing data. Data preprocessing is done to avoid over-fitting of the proposed model. 75% instances of the dataset are used for training whereas 25% data are used for testing.

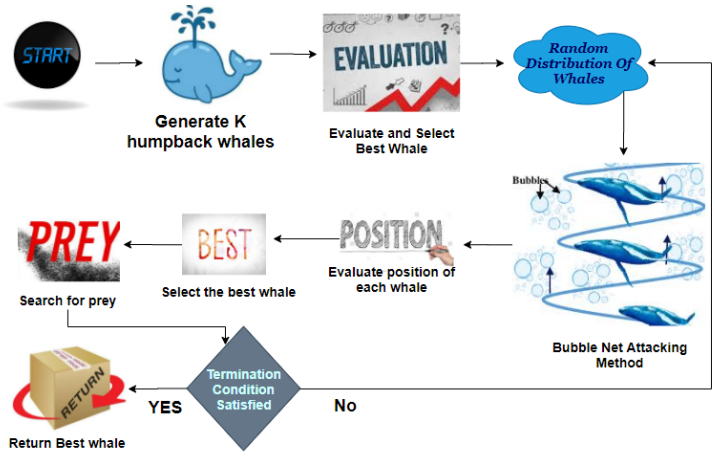


Fig. 2: Whale Optimization Algorithm Flow Chart

3.3 Feature Selection

WOA is used to perform feature selection in this phase. This step is done to find the best features among 30 attributes to feed the classification models. WOA is a meta-heuristic optimization proposed by Mirjalili & Lewis [24]. It is a nature-inspired approach that mimics the real-life behavior of a group of the largest mammals on the planet. WOA is a swarm-based technique that is designed based on the social behavior of humpback whales and takes inspiration from the bubble-net strategy unique to them for hunting in the ocean[15, 16]. Figure 2 shows the working flow chart of WOA. Humpback whales are the largest group of baleen whales and they usually spend their days as a group. They hunt small groups of krill and small fishes close to the surface by creating bubbles along a spiral path around their prey and then they swim up to the surface following this path [24, 19]. This hunting method is presented in Figure 3b .

Using this bubble-net hunting mechanism, Mirjalili & Lewis proposed a mathematical model and algorithm to solve optimization problems [12]. Working of WOA is depicted in Fig 2 The whale optimization algorithm consists of three main steps as discussed down below:

– **Encircling Prey:** To hunt, humpback whales can identify the location of their prey and attack them by encircling them. The whale optimization algorithm also works on a similar principal. As the best solution is not known beforehand, WOA assumes that the target prey is the current best solution found in the current iteration and that it is closest to the optimum solution. Once the current best solution or agent is found, other search agents also update their positions relevant to the selected solution encircling the prey solution. Equation 1,2,3 and 4 represent this behavior:

$$D = |CX(t) - X(t)| \tag{1}$$

$$X(t+1) = X(t) - AD \tag{2}$$

$$A = 2ar - a \tag{3}$$

$$C = 2r \tag{4}$$

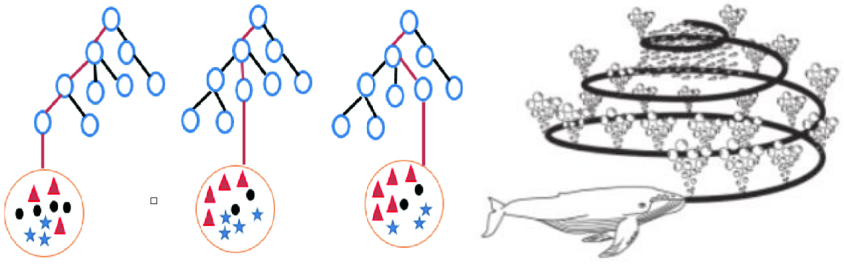
Here, t = the current iteration, A and C = coefficient vectors, X* = position vector of best solution which is updated if better solution is found, X = position vector, a = a number with value between 2 and 0, r = random vector [12].

- **Bubble-Net Attack:** This is the exploitation stage where humpback whales simultaneously move towards their prey in a shrinking circle and move in the spiral path simultaneously. WOA assumes that there is a 50% probability of a whale to select either of the following methods to catch their prey. This behavior of humpback whales can be described in the form of the following equations: Here, p is a random number with a value between 0 and 1 inclusive.
- **Search For Prey:** In addition to the bubble-net attack, the humpback whale also randomly searches for their prey. This is the exploration stage and the whale hunt their prey relevant to the position of other whales. To mimic this behavior, the value of 'A' vector is kept either less than -1 or greater than 1 to help WOA conduct a global search. This behavior is expressed in the form of the equation 5 and 6:

$$D = |CXrandX| \tag{5}$$

$$X(t+1) = Xrand - AD \tag{6}$$

Here, Xrand is a random position vector chosen from the current population [19].



(a) Extra Tree Classification [9]

(b) Hunting Behaviour of humpback whales [4]

Fig. 3: Extra Tree Classifier and Hunting Pattern of Whales

3.4 Classification Models

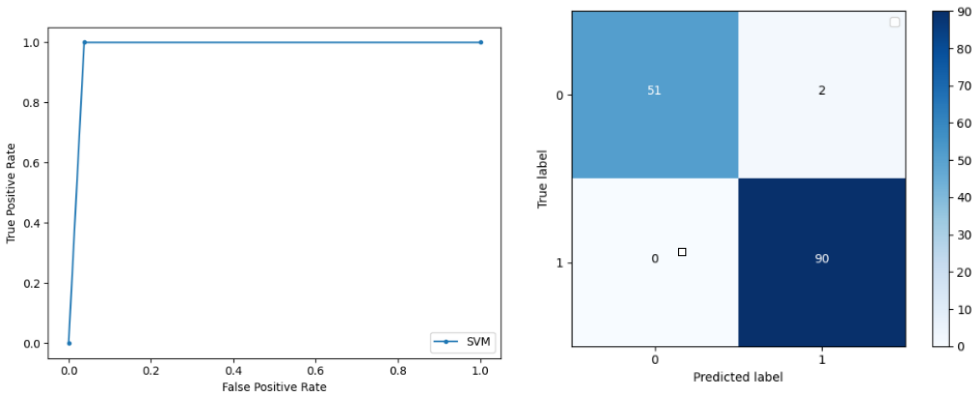
Classification model used in the proposed approach are machine learning algorithms and the basic purpose of using these classifiers is to predict the class label.The details of the classifiers used is discussed below:

- Extremely Randomized Tree Algorithm also known as Extra Tree,it is an ensemble of several decision trees. It is a forest of decision trees which is similar to the Random forest algorithm but differs from it in the way Decision trees build. To split decision tree's each node, every decision tree selects the best feature based on some selected criteria from a list of randomly selected K attribute. The Extra Trees algorithm creates unpruned trees and a large number of decision trees from training dataset.For regression and majority voting ,averaging the predictions is the technique used by this algorithm to make predictions of the decision trees[20].Figure 3a shows the classification by using Extremely Randomized Tree classifier.
- SVM performs classification for two-class problems. After the SVM model is trained using training data for each class, it can categorize new samples [2].it is a non-probabilistic binary linear classifier [20]. Support vector machines utilize the right hyper-plane to classify between two classes [20].There are many applications of SVM which include Protein Structure Prediction, Intrusion Detection, and Cancer Diagnosis[32].
- K-nearest neighbor, also known as K-NN, is an instance-based algorithm that classifies a sample based on the classes of its nearest neighbor. In this algorithm, K's closest examples are selected from the neighborhood of the sample that needs to be classified. A vote is taken among these examples and the new sample is assigned the most occurring class among its neighbors. A distance measure such as Euclidean distance is used to find the nearest neighbors of the sample [22]
- Naïve Bayes algorithm classifies samples using a probabilistic classification approach. It is a very simple and popular algorithm for classification problems. Gaussian Naïve Bayes is a variant of the Naive Bayes algorithm which is used for the classification of continuous data using the Gaussian distribution. In GNP, standard deviation and the mean of each given class corresponding to every sample in the training data is calculated and classification is performed according to these calculations [17]
- Stochastic Gradient Descent is an optimization technique that is used for training a classification model. Unlike gradient descent which utilizes all data samples to calculate the gradient of the cost function, SGD uses one randomly selected sample in each iteration. Although more noise is introduced in SGD due to random selection of samples, it is still much faster than GD to reach minima [32]
- Random Forest (RF) is an ensemble of tree-structured learning classifiers. It classifies a new sample based on the most occurring prediction made by these algorithms. Feature selection is used to grow the trees and at each node, random features are selected for splitting. This helps in minimizing over-fitting and as a result, RF classification is very fast [3]
- Logistic Regression, also known as parametric classification, utilizes "maximum likelihood estimation" for classification. It performs a probability analysis of the entire data to assign classes to new samples[13].
- Decision trees derive rules from the given training dataset and build a tree-like structure. The tree is grown by splitting nodes on the values of a feature. A criterion, such as information gain, is used to select the feature that best splits the tree and leads to a maximum decrease in entropy [13]

4 Experimental Analysis and Results

For experimentation, an optimal subset of features is given as input to the classification model. Wisconsin breast cancer dataset [] is used for the detection of breast Cancer. The dataset is splitted into 75% for training and 25% for testing. For the purpose of analyzing the performance of nine classifiers on WBDC dataset, the performance measures: accuracy, f-score, recall and precision to evaluate the efficiency of the proposed classification models.In this paper, a hybrid prediction algorithm is proposed for a fatal disease common in females called Breast cancer [8]. The drive behind this study is to provide a model of prediction using classifiers and feature selection for the early detection and prediction of breast cancer [30]. the Wisconsin dataset is used for prediction by using the best feature extracted through feature selection [14]. The best feature was predicted using WOA and then classification models were used for prediction. The proposed model is better than only using a classification algorithm for prediction because it reduces the number of features and is capable of handling large data. By using feature selection, Classification is performed easily. The WBDC dataset used is also normalized to reduce the amount of error. The proposed system achieved the highest accuracy of 99.30%, by using Extremely Randomized Tree. Other classifiers accuracy and evaluation measure is shown in Table 2

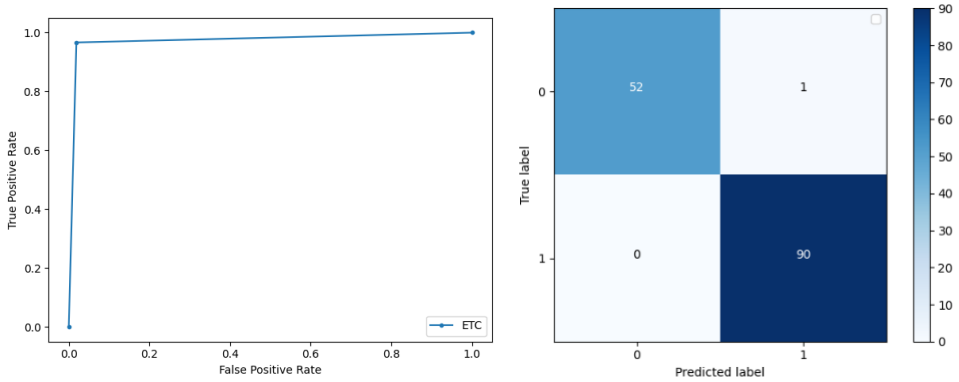
– **Support Vector Machine:** The support vector machine-classified the data with good accuracy. 98.60% accuracy is achieved by SVM with selected features extracted by WOA. The precision of both classes 0 and 1 is 1.00 and 0.98 respectively. Precision shows that the number of correctly identified positive results in both classes is more. F-score of the classification model for 0 and 1 class in SVM case is 0.98 and 0.99 respectively and recall for both classes is 0.96 and 1.00 respectively. These results depict that SVM can be used to predict breast cancer and early diagnosis. ROC and conclusion matrix of SVM classification is presented in Figure 4.



(a) Support Vector Machine ROC (b) Support Vector Machine confusion matrix

Fig. 4: Support Vector Machine ROC and confusion matrix

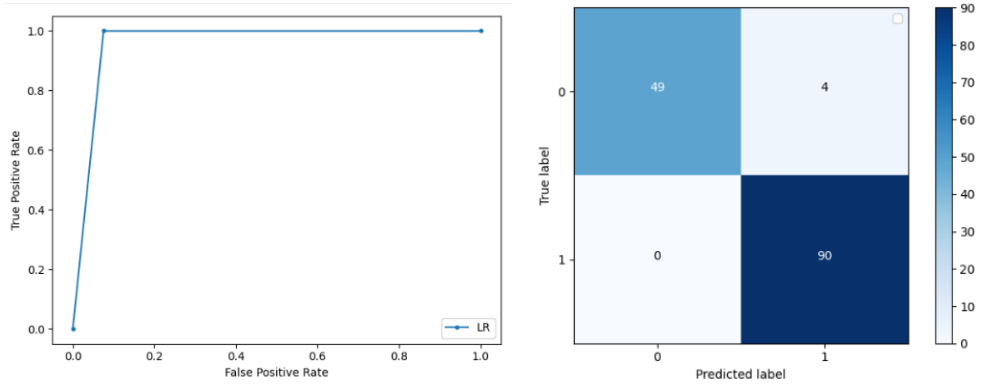
– **Extremely Randomized Trees Classifier:** EXT outperforms all the other classifiers and achieves accuracy of 99.30%. it aggregates the result of multiple co-related DTs and uses random sample for this purpose. The accuracy achieved shows that this classifier can be used to predict this fatal disease. The f-score of both 0 and 1 class in case of EXT is 1.00 and 0.99 and the recall for 0 and 1 was 0.98 and 1.00 respectively. Precision shows that correctly identified positive results are more and value for both classes is 0.99. ROC and conclusion matrix of EXT classification is presented in Figure 5.



(a) Extra Tree ROC (b) Confusion Matrix of Extra Tree

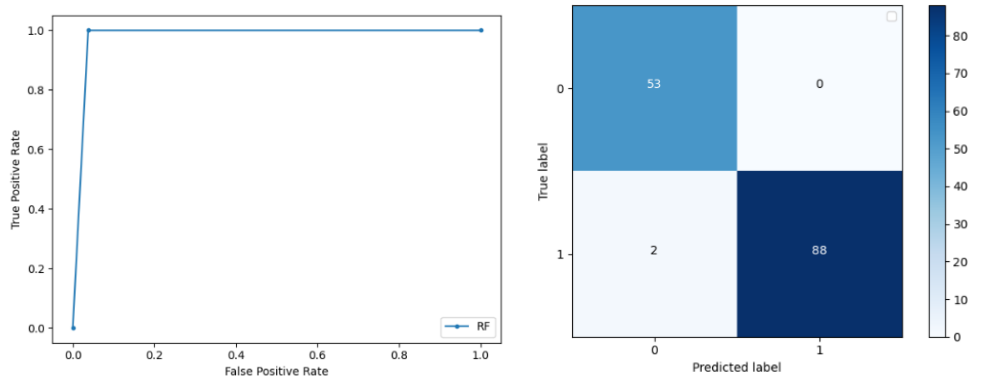
Fig. 5: Extra Tree ROC and confusion matrix

– **Logistic Regression:** Logistic regression classification was used to estimate the perimeters of logistic model and accuracy achieved by using LR classifier is 97.01% . The precision measure of regression for 0 and 1 class is 1.00 and 0.96 respectively and the f-score achieved is 0.96 and 0.98 respectively, the recall of the classifier is 0.98 and 0.96 respectively. ROC and conclusion matrix of LR classifier is presented in Figure 6.
– **Random Forest:** Random forest gives high accuracy when implemented on large dataset. The accuracy after applying this classifier on Wisconsin breast cancer dataset was 98.60% and the f-score calculated for 0 and 1 class is 0.98 and 0.99 respectively. The precision measure for RF was 0.96 and 1.00 for 0 and 1 class respectively. The recall of the classification model for class 0 and 1 was 1.00 and 0.98 respectively. These results shows that RF performs well on large dataset. ROC and conclusion matrix of RF classification is presented in Figure 7.
– **Kernal Support Vector Machine:** KSVM operates in high dimensions and does pattern analysis. The accuracy achieved by KSVM in the proposed solution was 94.41% and precision for class 0 and 1 was 0.98 and 0.93 respectively. The F-score for KSVM in both 0 and 1 class was 0.92 and 0.96 respectively. Recall for the KSVM classifier obtained was 0.87 and 0.99 for both 0 and 1 class respectively. ROC and conclusion matrix of KSVM classification is presented in Figure 8.



(a) Logistic Regression ROC (b) Confusion Matrix of Logistic Regression

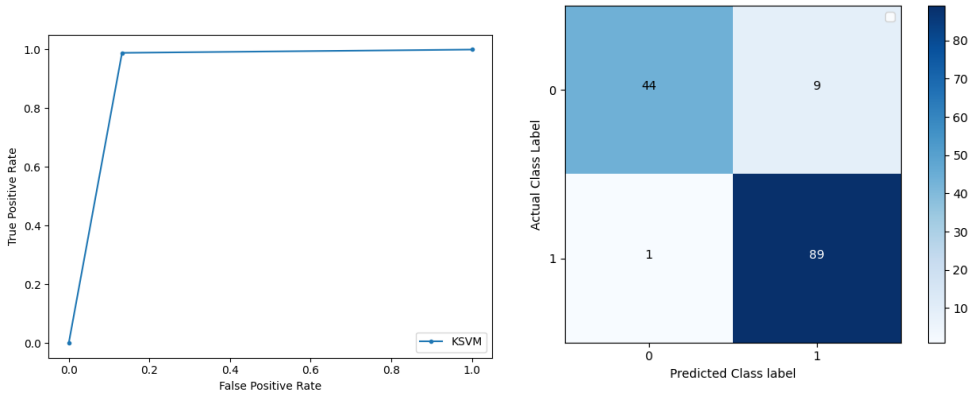
Fig. 6: Logistic Regression ROC and confusion matrix



(a) Random Forest ROC (b) Random Forest confusion matrix

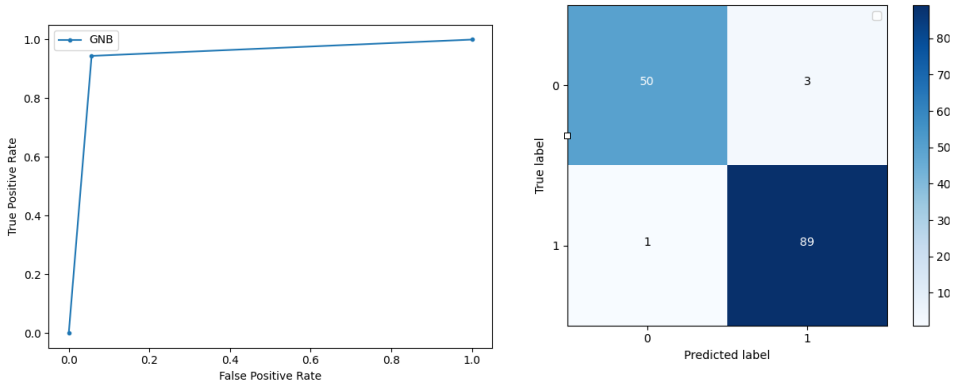
Fig. 7: Random Forest ROC and confusion matrix

– **Gaussian Naive Bayes:** GNB was used to find strong independence between the features of the dataset and it performs well when applied to the selected features of a huge dataset. The accuracy achieved by classifier is 94.40% and the precision measure of the 0 and 1 class is 0.91 and 0.97 respectively. The F-score achieved by using GNB is 0.93 and 0.96 for both 0 and 1 class respectively and the results shows that this classifier can be used for prediction of cancer but there are other classifiers which outperformed GNB in terms of accuracy. ROC and conclusion matrix of GNB classification is presented in Figure 9.
– **Stochastic Gradient Descent:** SGD performs well in large scale dataset but it is sensitive when it comes to feature scaling. The accuracy achieved by the classifier is 93.00% and the F-score for 0 and 1 class is 0.90 and 0.95 respectively. The precision and recall achieved for 0 and 1 class is 0.98, 0.91 and 0.83, 0.99 respectively. ROC and conclusion matrix of SGD classification is presented in Figure 10.



(a) Kernel Support vector Machine ROC (b) Kernel Support Vector Machine Confusion Matrix

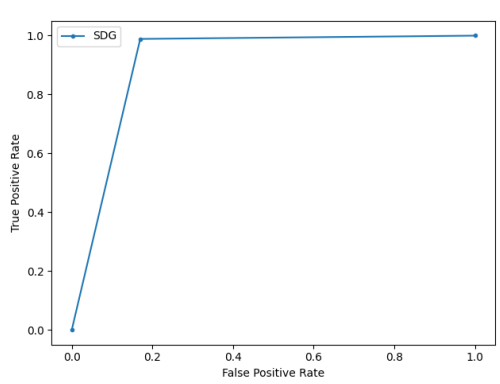
Fig. 8: Kernel Support Vector Machine ROC and confusion matrix



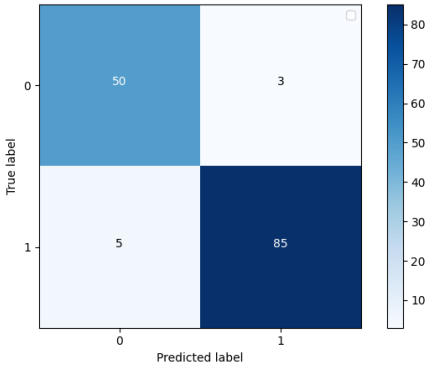
(a) Gaussian Naive Bayes ROC (b) Confusion Matrix of Gaussian Naive bayes

Fig. 9: Gaussian Naive Bayes ROC and confusion matrix

– **K-Nearest Neighbour:** K-NN relies on distance for classification and training data normalization can improve accuracy. The accuracy achieved by the K-NN used in proposed solution is 93.20% . F-score and precision of class 0 and 1 is 0.90, 0.95 and 0.92, 0.93 respectively.The results obtained by using K-Nearest Neighbour are adequate when there is no other classifier but it performed worse than the other classifiers used because of the value of the k selected on random basis. The recall of the system is 0.89 and 0.96. ROC and conclusion matrix of SGD classification is presented in Figure 11.
– **Decision Trees:** DT is used to split data according to a certain parameter. The accuracy achieved by the system is 97.20% and the F-score and precision of class 0 and 1 is 0.96, 0.98 and 0.95 and 0.99 respectively.The recall obtained is

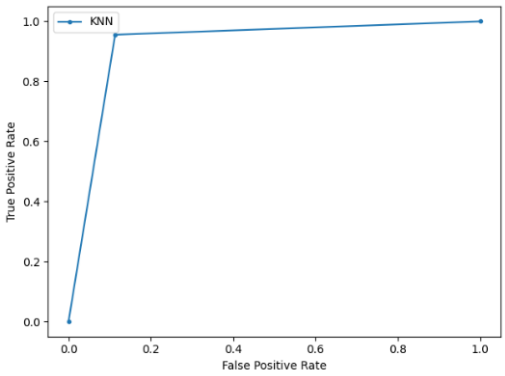


(a) Stochastic Gaussian Descent ROC

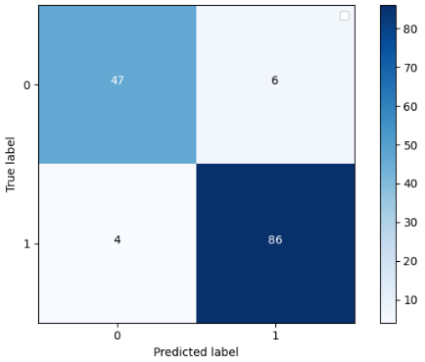


(b) Confusion Matrix of Stochastic Gradient Descent

Fig. 10: Stochastic Gaussian Descent ROC and confusion matrix



(a) K-nearest Neighbour ROC



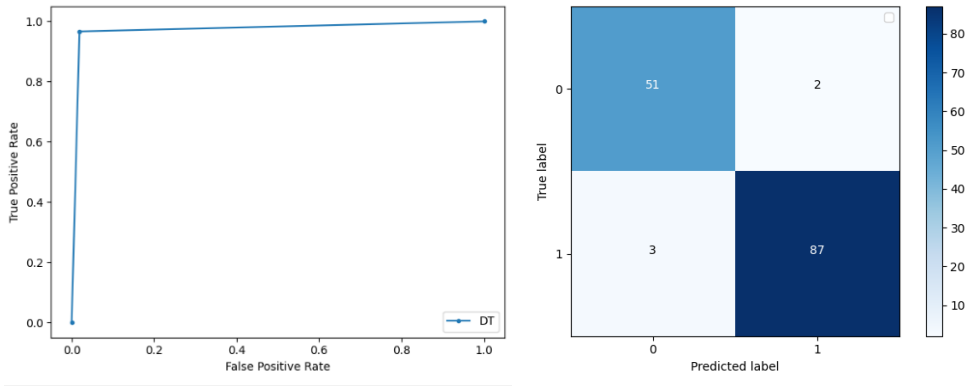
(b) Confusion Matrix of K- Nearest Neighbour

Fig. 11: K-Nearest Neighbour ROC and confusion matrix

0.98 and 0.97 for 0 and 1 class. ROC and conclusion matrix of DT classification is presented in Figure 12.

5 Discussion

There are a lot of randomization methods for trees and classification but building totally random trees is still far away.EXT works totally on random behaviour and, selects the cut points.at the end a randomized tree is build and then prediction and classification can occur by using that tree.The proposed approach is capable of accurately predicting whether a patient has breast cancer or not, even when limited data is available as it only chooses the best feature for classification. Medical



(a) Decision Tree ROC (b) Decision Tree Confusion Matrix

Fig. 12: Decision Tree ROC and confusion matrix

practitioners can easily use this model to check if a patient is at risk of getting breast cancer. Nowadays, medical facilities are getting expensive so this model can help predict patient diagnosis [31]. Table 2 shows the results obtained from proposed approach and extremely randomized tree classifier outperforms all other classifiers used for breast cancer prediction. SVM, EXT, DT, LR, and RF accuracy is higher as compared to other algorithms used. They can be used to predict and classify breast cancer to avoid the spreading of cancerous cells in the body. If the disease is diagnosed timely, patients can be saved from facing dire conditions and the treatment of this disease can be started on time. All classifiers used in this study performed well except SGD and K-NN. The reason behind this is because the K-NN classifier works on distance measures, the value of k also has an immense effect on the accuracy of the model. SDG performance lacks because of usage of the same learning rate for all parameters. All the other classifiers were able to classify the dataset with a high accuracy rate. ROC and confusion matrix of each classifier shows the performance of the classifier and EXT proves to be the best classifier among all the nine classifiers. Table 3 shows the result summary and comparison with the existing solutions. The ROC(Receiver Operating characteristics curve) and confusion matrix of each classifier is given.

Table 2: Result comparison Table

Classification Model	Accuracy	Precision	F-score	Recall
K-Nearest Neighbour	93.02%	0.92	0.9	0.89
		0.93	0.95	0.96
Decision Tree	97.20%	0.95	0.96	0.98
		0.99	0.98	0.97
Stochastic Gradient Descent	93.00%	0.98	0.90	0.83
		0.91	0.95	0.99
Kernel Support Vector Machine	94.41%	0.98	0.92	0.87
		0.93	0.96	0.99
Gaussian Naive Bayes	94.40%	0.91	0.93	0.94
		0.97	0.96	0.94
Random Forest	98.60%	0.96	0.98	1.00
		1.00	0.99	0.98
Logistic Regression	97.01%	1.00	0.96	0.92
		0.96	0.98	0.98
Support Vector Machine	98.60%	1.00	0.98	**0.96**
		0.98	0.99	1.00
Extra Tree	**99.30%**	**1.00**	**0.99**	**0.98**
		0.99	**0.99**	**1.00**

Table 3: Result summary

Authors	Dataset used	Algorithm	Results
Vikas et al. [10]	Breast Cancer medical dataset	Sequential Minimal Optimization, K-nearest Neighbor, Decision tree	Best results found by SMO, Accuracy:96.19%
Zheng et al[33]	Breast Cancer medical dataset	K-mean and Support vector machine	Accuracy:97.38%
Alghunaim et al[7]	Gene Expression	support vector machine, random forest, and decision trees	Support Vector Machine without feature selection outperformed with an accuracy of 99.68%,
Salama et al[28]	Wisconsin Breast cancer dataset	Support vector machine	SVM accuracy:96.99
Proposed Solution	**Wisconsin Breast Cancer dataset**	**SVM, K-NN, RF, LR, DT, GNB, SGD, ET, KSVM**	**Extra Tree outperformed other classifiers and achieved an accuracy of 99.30%**

6 Conclusion And Future Work

In this paper, a novel approach is proposed which utilizes Extremely randomized Tree and whale optimization algorithm for prediction of breast cancer. Decision tree, K-nearest neighbor, SGD, Random forest, Logistic regression, KSVM, Gaussian naive Bayes, and SVM are also used for the classification of the WBDC dataset. The results are then compared to choose the best classifier among the nine classification algorithms. WOA is used for feature selection and the extracted features are used to improve the classification accuracy. Feature selection is used to extract optimal features from the dataset to eliminate any unnecessary details. Pycharm platform and Wisconsin breast cancer diagnostic dataset was used for training and classification. The results showed that the proposed approach achieved the highest accuracy rate of 99.03% by using the Extremely Randomized

tree classifier. SVM also has an accuracy rate of 98.6%. Both EXT and SVM classifiers performed better than other algorithms. In the future, we intend to develop an application that shall let a user predict if the cancer is benign or malicious. This application would contribute positively to society and will help the medical community to detect cancer at early stages. Patients will also be able to use their medical reports to predict and analyze their chances of getting breast cancer and re occurrence of this disease.

References

1. Wisconsin breast cancer dataset. Accessed: 2020-04-3.
2. Muhammet Fatih Ak. A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications. In Healthcare, volume 8, page 111. Multidisciplinary Digital Publishing Institute, 2020.
3. Ozlem Akar and Oguz Gungor. Classification of multispectral images using random forest algorithm. Journal of Geodesy and Geoinformation, 1:105–112, 01 2012.
4. Ala Al-Zoubi, Hossam Faris, Ja'far Alqatawna, and Mohammad Hassonah. Evolving support vector machines using whale optimization algorithm for spam profiles detection on online social networks in different lingual contexts. Knowledge-Based Systems, 04 2018.
5. Ala'M Al-Zoubi, Ja'far Alqatawna, Hossam Faris, and Mohammad A Hassonah. Spam profiles detection on social networks using computational intelligence methods: The effect of the lingual context. Journal of Information Science, page 0165551519861599, 2019.
6. Al-Zoubi Ala'M, Hossam Faris, Ja'far Alqatawna, and Mohammad A Hassonah. Evolving support vector machines using whale optimization algorithm for spam profiles detection on online social networks in different lingual contexts. Knowledge-Based Systems, 153:91–104, 2018.
7. Sara Alghunaim and Heyam H Al-Baity. On the scalability of machine-learning algorithms for breast cancer prediction in big data context. IEEE Access, 7:91535–91546, 2019.
8. Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83:1064–1069, 2016.
9. Abdelkader Berrouachedi, Rakia Jaziri, and Gilles Bernard. Deep cascade of extra trees. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pages 117–129. Springer, 2019.
10. Vikas Chaurasia and S. Pal. A novel approach for breast cancer detection using data mining techniques. 2017.
11. Angeline Christobel and Y Sivaprakasam. An empirical comparison of data mining classification methods. International Journal of Computer Information Systems, 3(2):24–28, 2011.
12. Márcio Dias de Lima, Juliana de Oliveira Roque e Lima, and Rommel M Barbosa. Medical data set classification using a new feature selection algorithm combined with twin-bounded support vector machine. Medical & Biological Engineering & Computing, 58(3):519–528, 2020.

13. Stephan Dreiseitl and Lucila Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics, 35(5):352 – 359, 2002.
14. Ashutosh Kumar Dubey, Umesh Gupta, and Sonal Jain. Analysis of k-means clustering approach on the breast cancer wisconsin dataset. International journal of computer assisted radiology and surgery, 11(11):2033–2047, 2016.
15. Thippa Reddy Gadekallu, Dharmendra Singh Rajput, M Praveen Kumar Reddy, Kuruva Lakshmanna, Sweta Bhattacharya, Saurabh Singh, Alireza Jolfaei, and Mamoun Alazab. A novel pca–whale optimization-based deep neural network model for classification of tomato plant diseases using gpu. Journal of Real-Time Image Processing, pages 1–14, 2020.
16. Celestine Iwendi, Praveen Kumar Reddy Maddikunta, Thippa Reddy Gadekallu, Kuruva Lakshmanna, Ali Kashif Bashir, and Md Jalil Piran. A metaheuristic optimization approach for energy efficiency in the iot networks. Software: Practice and Experience, 2020.
17. H. Kamel, D. Abdulah, and J. M. Al-Tuwaijari. Cancer classification using gaussian naive bayes algorithm. In 2019 International Engineering Conference (IEC), pages 165–170, 2019.
18. Seyed Reza Kamel, Reyhaneh YaghoubZadeh, and Maryam Kheirabadi. Improving the performance of support-vector machine by selecting the best features by gray wolf algorithm to increase the accuracy of diagnosis of breast cancer. Journal of Big Data, 6(1):90, 2019.
19. Md Rezaul Karim, Oya Beyan, Achille Zappa, Ivan G Costa, Dietrich Rebholz-Schuhmann, Michael Cochez, and Stefan Decker. Deep learning-based clustering approaches for bioinformatics. Briefings in Bioinformatics, 2020.
20. Md Rezaul Karim, Ashiqur Rahman, João Bosco Jares, Stefan Decker, and Oya Beyan. A snapshot neural ensemble method for cancer-type prediction based on copy number variations. Neural Computing and Applications, pages 1–19, 2019.
21. Youness Khourdifi and Mohamed Bahaj. Applying best machine learning algorithms for breast cancer prediction and classification. In 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), pages 1–5. IEEE, 2018.
22. TP Latchoumi and Latha Parthiban. Abnormality detection using weighed particle swarm optimization and smooth support vector machine. 2017.
23. C Meera and D Nalini. Breast cancer prediction system using data mining methods. International Journal of Pure and Applied Mathematics, 119(12):10901–10911, 2018.
24. Seyedali Mirjalili and Andrew Lewis. The whale optimization algorithm. Advances in engineering software, 95:51–67, 2016.
25. Hardi M Mohammed, Shahla U Umar, and Tarik A Rashid. A systematic and meta-analysis survey of whale optimization algorithm. Computational intelligence and neuroscience, 2019, 2019.
26. G Thippa Reddy, M Praveen Kumar Reddy, Kuruva Lakshmanna, Rajesh Kaluri, Dharmendra Singh Rajput, Gautam Srivastava, and Thar Baker. Analysis of dimensionality reduction techniques on big data. IEEE Access, 8:54776–54788, 2020.
27. Sapiah Binti Sakri, Nuraini Binti Abdul Rashid, and Zuhaira Muhammad Zain. Particle swarm optimization feature selection for breast cancer recur-

rence prediction. IEEE Access, 6:29637–29647, 2018.
28. Gouda I Salama, M Abdelhalim, and Magdy Abd-elghany Zeid. Breast cancer diagnosis on three different datasets using multi-classifiers. Breast Cancer (WDBC), 32(569):2, 2012.
29. Gehad Ismail Sayed, Ashraf Darwish, Aboul Ella Hassanien, and Jeng-Shyang Pan. Breast cancer diagnosis approach based on meta-heuristic optimization algorithm inspired by the bubble-net hunting strategy of whales. In International Conference on Genetic and Evolutionary Computing, pages 306–313. Springer, 2016.
30. Ali Sharifi and Kamal Alizadeh. Prediction of breast tumor malignancy using neural network and whale optimization algorithms (woa). 2019.
31. K Sivakami and Nadar Saraswathi. Mining big data: breast cancer prediction using dt-svm hybrid model. International Journal of Scientific Engineering and Applied Science (IJSEAS), 1(5):418–429, 2015.
32. Susana M Vieira, Luís F Mendonça, Goncalo J Farinha, and João MC Sousa. Modified binary pso for feature selection using svm applied to mortality prediction of septic patients. Applied Soft Computing, 13(8):3494–3504, 2013.
33. Bichen Zheng, Sang Won Yoon, and Sarah Lam. Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. Expert Systems with Applications, 41:1476–1482, 08 2013.