Research Article

Journal of Scientific Enquiry
Vol: 2, Pages: 35-41

Application of Particle Swarm Optimization on Wisconsin Diagnosis Breast Cancer Dataset

Anupam Sen

Department of Computer Science, Government General Degree College Singur, Hooghly, India

Received: 28.10.2021; Accepted: 13.12.2021; Published Online: 30.06.2022

Abstract:

Machine Learning techniques are playing an important role within the medical field. Machine learning algorithms can be applied to develop models for better prediction of breast cancer. Usually, a medical dataset contains a large set of features. Classification accuracy can be increased by selecting appropriate features. This paper aims to increase the accuracy of the existing few data mining algorithms. It embeds Particle swarm intelligence for selecting features from Wisconsin Diagnosis Breast Cancer Dataset (WDBC). Three renowned classifiers Decision Stump, J48 pruned tree and Naive Bayes are used to improve accuracy, Kappa statistic, Mathew's Correlation Coefficient, Precision, F-measure, Recall, Mean Absolute Error (MAE), Root Mean Square Error (RMSE). This approach can be further embedded into IoT based breast cancer prediction support systems. Proposed method can be helpful for the medical expert to diagnose breast cancer competently.

Keywords:

Particle swarm optimization, Decision Stump, Kappa statistic, Mean Absolute Error (MAE), Root Mean Square Error (RMSE).

1. Introduction:

World Health Organization (WHO) reports document that more than 2.1 million deaths are occurring worldwide due to breast cancer which is the most predominant cancer among women[1] . Presently, to detect early-stage breast cancer X-ray mammography is used. In an asymptomatic population this method is very useful to detect breast cancer in a systematic way. In its initial phase, mammographic images are employed to differentiate small masses and micro-calcifications to spot breast cancer [2] . Generally, in case of breast cancer when the patient becomes aware of their condition the chances of survival become bleak because of absence of pain sensations or symptoms in the early stages. It is true that accurate and timely diagnosis can greatly increase survival chances and aid in reducing treatment costs. Today modern diagnosis involves precise evaluation of patient data and expert decision coupled with varied machine learning approaches and pattern recognition which are proposed to provide supportive aid to experts in their decision- making process. The dominant role of these approaches is in extraction of informative knowledge about patient's data and reduction in the time and cost of diagnosis. In this context, many methods have been suggested based on the Swarm Intelligence approach. As a discipline of artificial intelligence, Swarm intelligence (SI) deals with the designing of intelligent multi- agent systems, drawing knowledge from the collective behaviour of ants, termites, bees and wasps which are called social insects as well as drawing references from other herd animal communities such as birds or fish. SI approach thereby utilizes intelligent agents for exploring the problem space and extracting perfecting solutions. Of late, SI is being used in optimizing problems, which most dominantly involves the medical field such as diagnosis, prediction, treatment and screening [3].

2. RELATED WORKS

L. Gao, M. Ye, and C. Wu have experimented on 9 cancer datasets using SVM (Support Vector Machine) which was optimized by PSO (Particle Swarm Optimization) together with ABC (Artificial Bee Colony) that forecasted breast cancer among patients [4]. Swesi, I. M. A. O and Bakar proposed a feature clustering algorithm on high dimensional data for improving the accuracy of the approach, and lowering the computational cost [5]. T. Advancements et al. have carried out research work on Elitism Particle Swarm Optimization (EPSO) and Recursive Feature Reduction (RFR) for selecting genes to generate improved classifying performance which could be biologically relevant to cancer [6]. M. R. Mohebian et al. [4] have carried out research work and have built a Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) using Optimized Ensemble Learning to identify breast cancer recurrence within 5 years after diagnosis [7]. D. A. Utami and Z. Rustam presented a comparison of PSO and SVM and ABC and SVM machine

[1] Email: rabi.pandey@gmail.com

DOI: https://doi.org/10.54280/jse.222106

learning algorithms for identification of indicators of breast cancer. ABC–SVM method showed improved performance with a precision rate of 88% cancer classification when contrasted with PSO–SVM method, having a precision rate of 87 % [8]. S. B. Sakri et al. have implemented PSO using three well-approved classifying algorithms, naïve Bayes, IBK, and REPTree, having with and without feature selection algorithm for breast cancer prediction [9]. M. Mahajan et al. proposed a particle swarm optimization method to enhance the performance of the kNN classifier [10]. S. Jeyasingh and M. Veluchamy proposed Modified Bat Algorithm (MBA) for feature optimization with Random Forest (RF) classifier. [11]. M. A. Rahman and R. C. Muniyandi implemented a selection method which had a two-step character based on Artificial Neural Networks using 15 Neurons used to increase the performance of the classifier [12]. B. Al-Shargabi et al. implemented Multilayer perceptron with feature selection to predict breast cancer which obtained an accuracy rate of 97.70% [13]. S. Pravesjit et al. proposed a hybrid PSO with ROA algorithm for breast cancer prediction on Wisconsion Breast Cancer Dataset which obtained an accuracy 98% [14].

3. PROPOSED METHODOLOGY

In the proposed work, Wisconsin Diagnosis Breast Cancer (WDBC) balanced dataset is obtained from the UCI ML Repository [15]. Dataset has 569 instances with 31 features and a class variable, i.e. (M = malignant, B = benign). All information about attributes is presented in Table 1. Particle Swarm Optimization (PSO) with the best first method is used for selecting best fit attributes from the dataset. Table. 2 contains the features selected by the PSO search method. Proposed layout of the model is depicted in fig. 1.

Table. 1. Attribute information

(1) ID number
(2) Diagnosis M=malignant, B=benign
(3–32) Ten real-valued features are computed for each cell nucleus:
(a) radius (mean of distances from 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 (perimeter^2/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)

. Table.2. Selected attribute by PSO method

SL No.	Attribute name
1.	texture_mean
2.	area_mean
3.	concavity_mean
4.	concave points_mean
5.	area_se
6.	symmetry_se
7.	perimeter_worst
8.	area_worst
9.	smoothness_worst
10.	concavity_worst
11.	texture_mean

Table 3. Accuracy, kappa statistic, MAE, RMSE without PSO search

Classification Algorithm	Accuracy	Kappa statistic	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)
Decision Stump	89.10%	0.7615	0.1679	0.3108
J48 pruned tree	94.02%	0.8732	0.0665	0.241
Naive Bayes	94.20%	0.8751	0.0574	0.2225

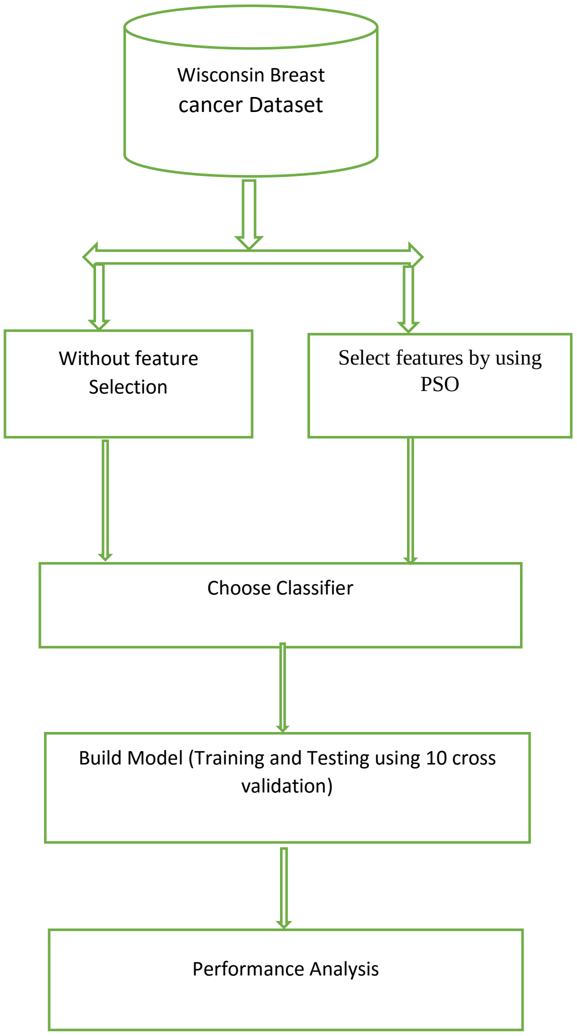

Wisconsin Breast cancer Dataset
Without feature Selection
Select features by using PSO
Choose Classifier
Build Model (Training and Testing using 10 cross validation)
Performance Analysis

Fig. 1. Layout of proposed model

Table 4. Accuracy, kappa statistic, MAE, RMSE with PSO search

Classification Algorithm	Accuracy	Kappa statistic	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)
Decision Stump	88.92 %	0.7559	0.1692	0.3138
J48 pruned tree	93.32 %	0.8582	0.0723	0.2544
Naive Bayes	92.97 %	0.8491	0.07	0.2585

Table 5. Performance Analysis of the all models without feature selection

Classification Algorithm	Precision	Recall	F-Measure	MCC	ROC Area	Class
Decision Stump	0.912	0.778	0.840	0.761	0.874	M
	0.879	0.955	0.915	0.761	0.874	B
J48 pruned tree	0.895	0.929	0.912	0.859	0.931	M
	0.957	0.936	0.946	0.859	0.931	B
Naive Bayes	0.913	0.896	0.905	0.849	0.980	M
	0.939	0.950	0.944	0.849	0.980	B

Table 6. Performance Analysis of the all models with feature selection

Classification Algorithm	Precision	Recall	F-Measure	MCC	ROC Area	Class
Decision Stump	0.899	0.797	0.797	0.765	0.885	M
	0.887	0.947	0.916	0.765	0.885	B
J48 pruned tree	0.905	0.939	0.921	0.874	0.936	M
	0.963	0.941	0.952	0.874	0.936	B
Naive Bayes	0.937	0.906	0.921	0.875	0.986	M
	0.945	0.964	0.954	0.875	0.986	B

Table 7. Correct and incorrect classification of cancer without feature selection

Classification Algorithm	Malignant		Benign	
	Correct	Incorrect	Correct	Incorrect
Decision Stump	165	47	341	16
J48 pruned tree	197	15	334	23
Naive Bayes	190	22	339	18

Table 8. Correct and incorrect classification of cancer with feature selection

Classification Algorithm	Malignant		Benign	
	Correct	Incorrect	Correct	Incorrect
Decision Stump	169	43	338	19
J48 pruned tree	199	13	336	21
Naive Bayes	192	20	344	17

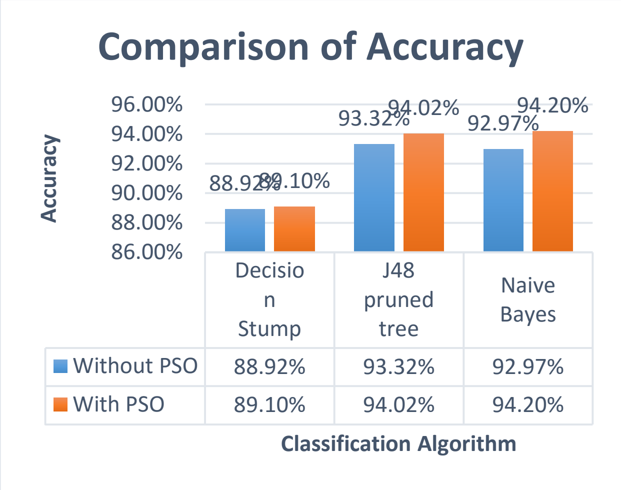


Fig. 2. Comparison of Accuracy without PSO and with PSO

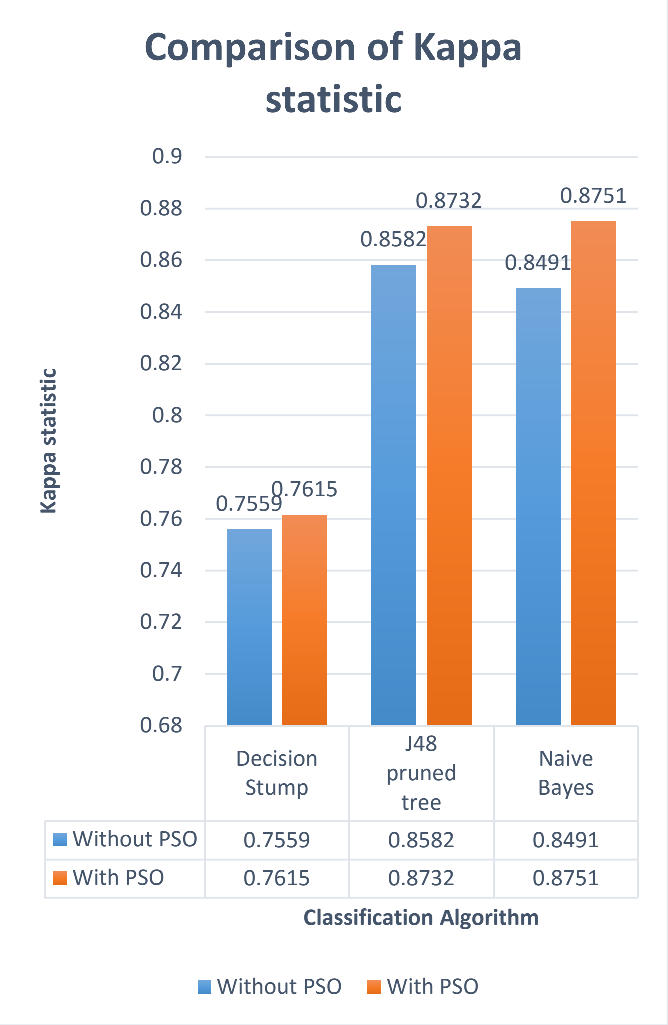


Fig. 3. Comparison of Kappa Statistic without PSO and with PSO

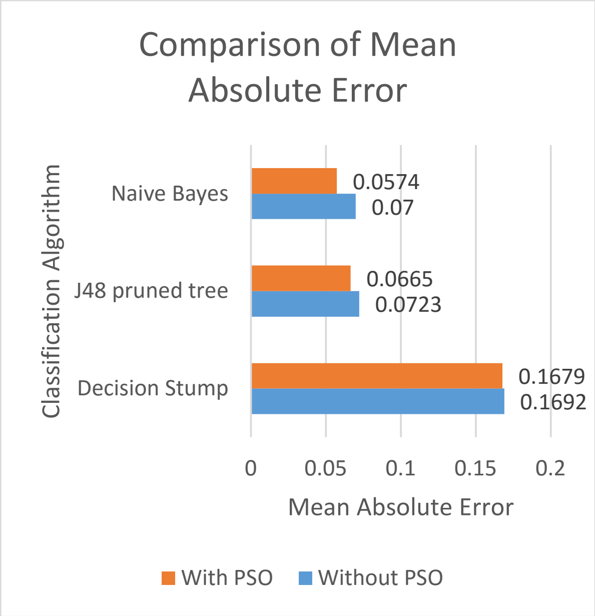


Fig. 4. Comparison of MAE without PSO and with PSO

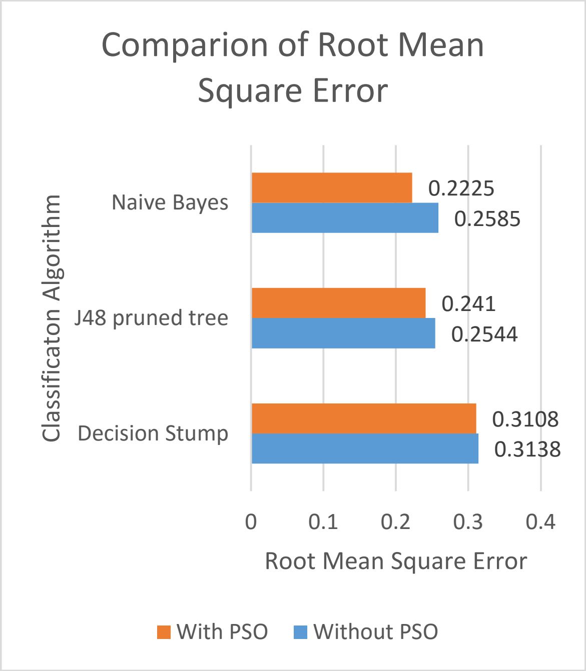
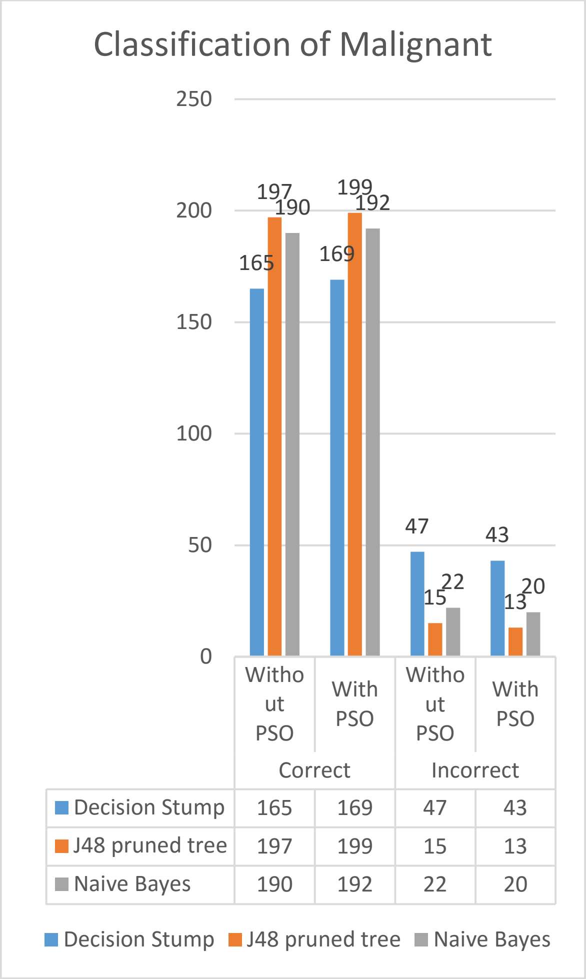
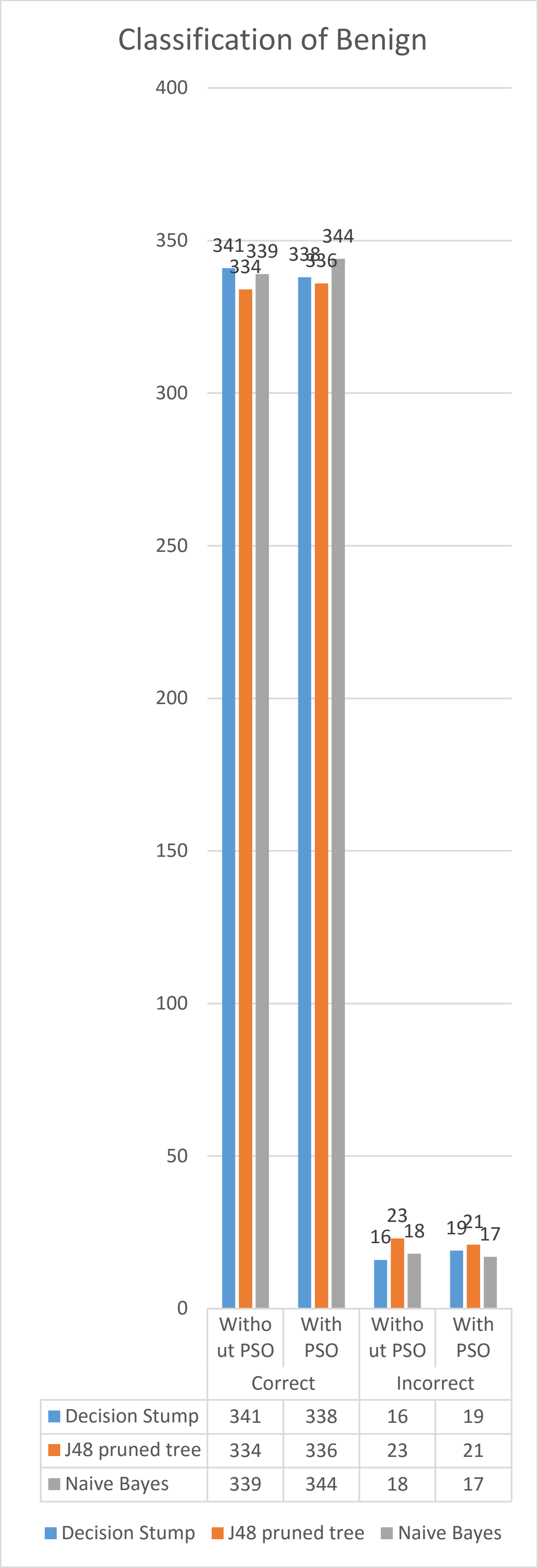


Fig. 5. Comparison of RMSE without PSO and with PSO



	Correct		Incorrect	
	Without PSO	With PSO	Without PSO	With PSO
Decision Stump	165	169	47	43
J48 pruned tree	197	199	15	13
Naive Bayes	190	192	22	20

Fig. 6. Classification of malignant by all the models with PSO and without PSO



	Correct		Incorrect	
	Without PSO	With PSO	Without PSO	With PSO
Decision Stump	341	338	16	19
J48 pruned tree	334	336	23	21
Naive Bayes	339	344	18	17

Fig. 7. Classification of benign by all the models with PSO and without PSO

4. EXPERIMENTAL RESULTS

For this work, the dataset is taken from UCI repository. WEKA (The Waikato Environment for Knowledge Analysis) software was employed to run machine learning techniques. First, three classifiers Decision Stump, J48 pruned tree and Naive Bayes were used to construct the model without feature selection. Then, PSO with the best first method was used for selecting best fit attributes from the dataset. Three classifiers Decision Stump, J48 pruned tree and Naive Bayes was applied on the selected features to construct the model. Here the 10-fold cross validation method was employed for training, testing and validation purposes. Table 3 represents the information about accuracy, Kappa statistic, Mean Absolute Error (MAE), Root Mean Square Error (RMSE) of the different classifier algorithms without selecting features. Table 4 represents information about accuracy, Kappa statistic, Mean Absolute Error (MAE), Root Mean Square Error (RMSE) of the different classifier algorithms based on attribute selection method for selecting features. Table 5 and 6 shows the performance analysis of all models without feature selection and with feature selection respectively. Table 7 represents correct and incorrect classification of malignant, benign of different classification algorithms without feature selection. Table 8 represents correct and incorrect classification of malignant, benign of different classification algorithms with feature selection. The bar graph in figure 2 and figure 3 gives the comparison of accuracy and kappa statistic obtained without PSO and with PSO of different classification algorithms. Breast cancer prediction accuracy and kappa statistic of Decision Stump, J48 pruned tree and Naive Bayes algorithms performance is enhanced when PSO is applied. Accuracy and kappa statistic of Naive Bayes classifier is better than other classifiers. Mean Absolute Error, Root Mean square Error without PSO and with PSO of different classification algorithm is shown in fig. 4, fig.5 respectively. The figure 4 shows the Mean Absolute Error is minimized for Naive Bayes classifier with PSO. The figure 5 shows the Root Mean Square Error is minimized for Naive Bayes classifier with PSO. Fig 6 depicts correct and incorrect instances of malignant by all the classifiers with PSO and without PSO. Fig 7 represents correct and incorrect instances of benign by all the classifiers with PSO and without PSO.

5. CONCLUSION AND FUTURE WORK

The intention of the work was to enhance the performance of the different classifiers so that they can more accurately identify the early diagnosis of breast cancer. The value in Table 3 compares accuracy, kappa statistic, Mean Absolute Error, Root Mean Square of different algorithms based on without feature selection. The accuracy for J48 pruned tree classification algorithm is 93.32% which is seen to perform than other algorithms with also higher kappa statistic and minimum Root Square Mean Error but the Mean Absolute Error is found to be least for Naive Bayes. The value in Table 4 compares accuracy, kappa statistic, Mean Absolute Error, Root Mean Square Error of different classification algorithms based on the PSO attribute selection method. In this case the accuracy for Naive Bayes classification algorithm is 94.20% which performs better than other algorithms with higher Kappa statistic and the Mean Absolute Error and Root Mean Square Error was minimum for Naive Bayes classifier. All three model's performance is better with the PSO feature selection method compared to without the PSO feature selection method. Drawing from the findings it can be concluded that feature extraction and machine learning algorithms play an essential role in identifying the early diagnosis of breast cancer to reduce cost and time. The future work of the research work is to improve the accuracy of breast cancer prediction by applying newer algorithms and various feature selection methods. Breast cancer prediction can be automated using real time data. Early diagnosis and reduced cost can improve healthcare facilities in future.

REFERENCES

[1] Y. D. Austria, M. L. Goh, L. Sta. Maria Jr., J.-A. Lalata, J. E. Goh, and H. Vicente, "Comparison of Machine Learning Algorithms in Breast Cancer Prediction Using the Coimbra Dataset," Int. J. Simul. Syst. Sci. Technol., pp. 1–8, 2019, doi: 10.5013/ijssst.a.20.s2.23.

[2] P. Mora et al., "Improvement of early detection of breast cancer through collaborative multi-country efforts: Medical physics component," Phys. Medica, vol. 48, no. March, pp. 127–134, 2018, doi: 10.1016/j.ejmp.2017.12.021.

[3] H. Zamani and M.-H. Nadimi-Shahraki, "Swarm Intelligence Approach for Breast Cancer

Diagnosis," Int. J. Comput. Appl., vol. 151, no. 1, pp. 40–44, 2016, doi: 10.5120/ijca2016911667.

[4] L. Gao, M. Ye, and C. Wu, "Cancer classification based on support vector machine optimized by particle swarm optimization and artificial bee colony," Molecules, vol. 22, no. 12, 2017, doi: 10.3390/molecules22122086.

[5] C. Technology, "How to cite this article: Swesi, I. M. A. O., & Bakar, A. A. (2019). Feature clustering for pso-based feature construction on high-dimensional data.," vol. 4, no. 4, pp. 439–472, 2019.

[6] T. Advancements, R. Nagpal, and R. Shrivas, "Cancer Classification Using Elitism PSO Based Lezy IBK on Gene Expression Data," vol. 1, no. 4, pp. 19–23, 2015.

[7] M. R. Mohebian, H. R. Marateb, M. Mansourian, M. A. Mañanas, and F. Mokarian, "A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) Using Optimized Ensemble Learning," Comput. Struct. Biotechnol. J., vol. 15, pp. 75–85, 2017, doi: 10.1016/j.csbj.2016.11.004.

[8] D. A. Utami and Z. Rustam, "Gene selection in cancer classification using hybrid method based on Particle Swarm Optimization (PSO), Artificial Bee Colony (ABC) feature selection and support vector machine," AIP Conf. Proc., vol. 2168, 2019, doi: 10.1063/1.5132474.

[9] S. B. Sakri, N. B. Abdul Rashid, and Z. Muhammad Zain, "Particle Swarm Optimization Feature Selection for Breast Cancer Recurrence Prediction," IEEE Access, vol. 6, pp. 29637–29647, 2018, doi: 10.1109/ACCESS.2018.2843443.

[10] M. Mahajan, S. Kumar, B. Pant, K. Joshi, and V. Tripathi, "PSO Optimized Nearest Neighbor Algorithm," Int. J. Eng. Adv. Technol., vol. 9, no. 2, pp. 1508–1513, 2019, doi: 10.35940/ijeat.b3574.129219.

[11] S. Jeyasingh and M. Veluchamy, "Modified bat algorithm for feature selection with the Wisconsin Diagnosis Breast Cancer (WDBC) dataset," Asian Pacific J. Cancer Prev., vol. 18, no. 5, pp. 1257–1264, 2017, doi: 10.22034/APJCP.2017.18.5.1257.

[12] M. A. Rahman and R. C. Muniyandi, "An enhancement in cancer classification accuracy using a two-step feature selection method based on artificial neural networks with 15 neurons," Symmetry (Basel)., vol.12, no.2,2020, doi: 10.3390/sym12020271.

[13] B. Al-Shargabi, F. Al-Shami, and R. S. Alkhawaldeh, "Enhancing Multi-Layer Perceptron for Breast Cancer Prediction," Int. J. Adv. Sci. Technol., vol. 130, no. September, pp. 11–20, 2019, doi: 10.33832/ijast.2019.130.02.

[14] S. Pravesjit, P. Longpradit, K. Kantawong, R. Pengchata and N. Oul, "A Hybrid PSO with Rao Algorithm for Classification of Wisconsin Breast Cancer Dataset," 2021 2nd International Conference on Big Data Analytics and Practices (IBDAP), 2021, pp. 68-71, doi: 10.1109/IBDAP52511.2021.9552152.

[15] UCI "Machine Learning Repository" https://archive.ics.uci.edu/ml/index.php.