Performance Analysis Using Data Mining on Wisconsin Diagnostic Breast Cancer Dataset

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

***Abstract*— Machine Learning techniques (ML) are playing a pivotal role in the medical field. Early diagnosis is required to prevent breast cancer. In this research Wisconsin Diagnostic Breast Cancer (WDBC) dataset with 32 predictors are analyzed to classify breast cancer. In this study attribute selection method CfsSubsetEval evaluator with best first search method is used for feature selection. Various machine learning algorithms are applied to compare Accuracy, kappa statistics, mean absolute error and root mean square error on selected features of this dataset to determine how prevalent those predictors in determining breast cancer. How the feature selection plays an important role to improve the performance of various classifier algorithms.**

Keywords—Decision Stump, kappa statistics, Random Forest, Logistic Regression, MAE, RMSE.

I. Introduction

According to World Health origination breast cancer is the most predominant cancer among women. Maximum number of cancer-related deaths among women were reported due to breast cancer causing 2.1 million death each year [1]. To detect early stage breast cancer X-ray mammography is used at present. In asymptomatic population this method is very useful to detect breast cancer in a systematic way. Mammography images are used to differentiate small masses and micro-calcifications to spot breast cancer in its starting phase [2]. At present, mammography is widely used standard screening process for breast cancer. In Breast cancer prediction incorrect classifications of mammograms can be improved. Still a challenge to develop a cheap and easily accessible method from those predictors. Biomarkers present in the blood samples may provide alternative ways to better diagnose breast cancer among women [3].

II. Background of the study

Good outcome in treatment can be achieved by early diagnosing of breast cancer. More screening tools are required for healthy predictive models based on data which may be collected in blood analysis and routine consultation. Through routine blood analysis like Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP-1, Age and Body Mass Index (BMI) can be collected. In this work, try to assess how models based on data may be used to forecast the presence of breast cancer. These parameters are also related to obesity-associated breast cancer, [4]. Wisconsin breast cancer diagnosis (WBCD) dataset has been widely used. In this paper, various machine learning algorithms applied on breast cancer diagnosis and prognosis were discussed [5]. In another study shows Metabolic Syndrome, specifically insulin resistance and abdominal fat women after menopause have a large possibility of breast cancer. Subclinical insulin resistance, Homeostasis Model Assessment – Insulin Resistance (HOMA-IR) can be used to identify patients. For high risk patients this is important for prevention and testing [6]. In this study, Random forest and Naive Bayes were used as feature selection method and rank the feature importance [7]. In this paper [8], classifier model Deep Neural network (DNN) and recursive feature elimination (RFE) for feature selection were used to obtain 98.62% accuracy. In this work, the optimal activation function is used to reduce the classification error by using fewer blocks. In this article, the combination of age, body mass index (BMI), and metabolic parameters was determined as a potential reasonable and effective predictor for breast cancer [9].

III. Material and Methods

Many different techniques were used for the detection of breast cancer when related works were analyzed. There are several datasets available for the detection of breast cancer. In this paper, we use Wisconsin Diagnostic Breast Cancer (WDBC) dataset with 569 instances and 32 attributes taken from UCI ML Repository [10]. Dr. William H. Wolberg of the University of Wisconsin created this dataset to diagnose breast cancer, i.e., (M = malignant, B = benign). Research methodology is shown in fig1. Table I contains the features selected by CfsSubsetEval evaluator with best first search method.

Fig. 1 Research methodology

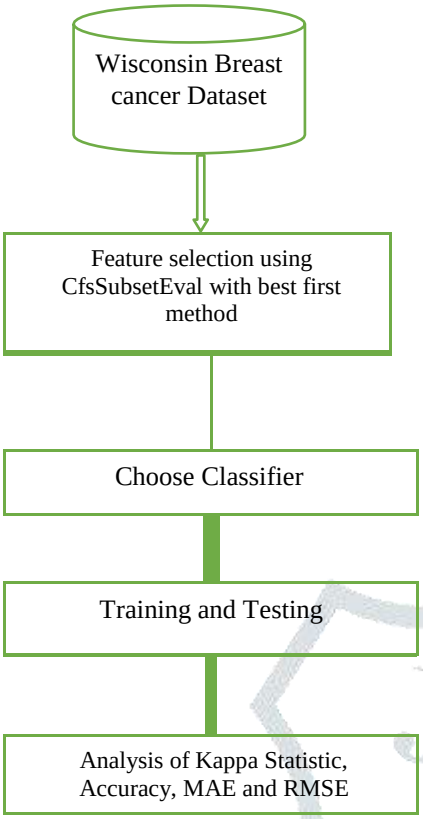

TABLE I. SELECTED ATTRIBUTE BY CFSSUBSETEVAL METHOD

Selected Attribute
texture_mean
concavity_mean
concave points_mean
area_se
symmetry_se
radius_worst
perimeter_worst
area_worst
smoothness_worst
concavity_worst
concave points_worst

TABLE II. ACCURACY, KAPPA STATISTIC, MAE, RMSE BASED ON WITHOUT FEATURE SELECTION (WFS).

Classification Algorithm	Accuracy	Kappa statistic	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)
DS	88.92 %	0.7559	0.1692	0.3138
J48 pruned tree	93.32%	0.8582	0.0723	0.2544
LMT	97.18 %	0.9395	0.0435	0.1409
RF	96.48 %	0.9242	0.0752	0.1717
SMO	97.71%	0.9507	0.0228	0.1512
Logitboost	96.83 %	0.9321	0.062	0.1774
IBK(K=5)	97.36 %	0.943	0.046	0.1532
MEF	94.72 %	0.8866	0.0527	0.2296
MLP	97.36 %	0.9431	0.0321	0.1577
LR	97.18 %	0.9395	0.0435	0.1409

TABLE III ACCURACY, KAPPA STATISTIC, MAE, RMSE BASED ON FEATURE SELECTION (FS).

Classification Algorithm	Accuracy	Kappa statistic	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)
DS	88.92%	0.7559	0.1692	0.3138
J48 pruned tree	94.02 %	0.8732	0.067	0.2413
LMT	97.53 %	0.9471	0.0487	0.153
RF	96.13 %	0.9168	0.0697	0.1766
SMO	97.01 %	0.9354	0.0299	0.1728
Logitboost	94.72 %	0.8868	0.0639	0.1881
IBK(K=5)	97.53 %	0.9473	0.0492	0.1641
MEF	97.18 %	0.9396	0.0281	0.1677
MLP	97.18 %	0.9395	0.0342	0.1599
LR	97.53 %	0.9471	0.0487	0.153

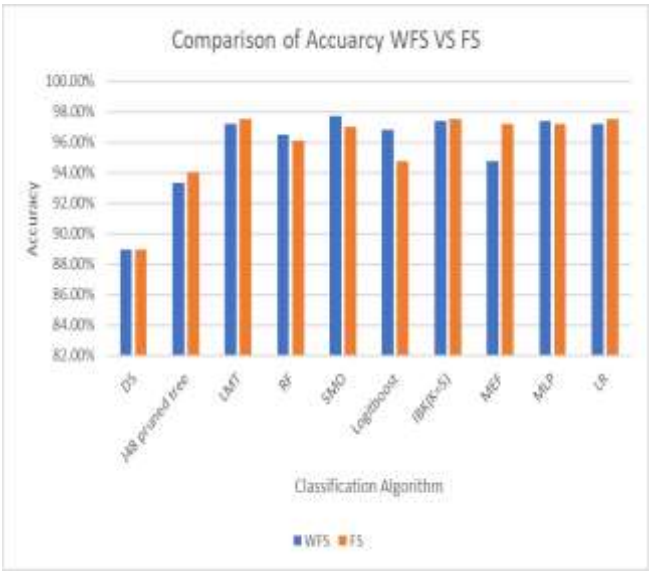

Fig. 2. Comparison of Accuracy based on WFS and FS

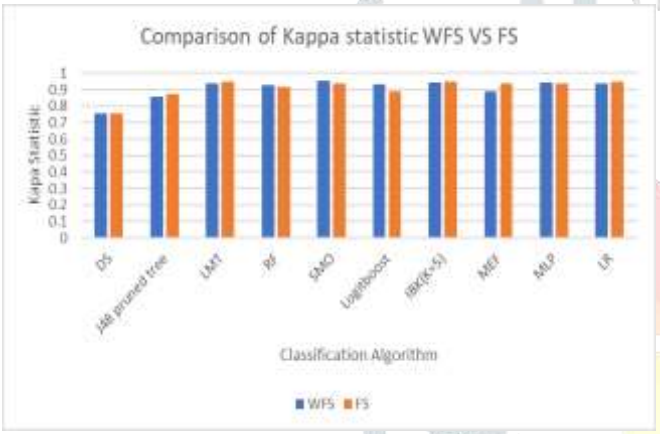

Fig. 3. Comparison of Kappa statistic based on WFS and FS.

IV. Application and Results

In this paper dataset from UCI repository is used [10]. Then apply methods for feature selection and then apply Machine learning algorithms to predict accuracy and statistics value. WEKA (The Waikato Environment for Knowledge Analysis) software is used for machine learning techniques. In this paper attribute selection method CfsSubsetEval is used for selecting features. Different MLR classification algorithms are applied through WEKA such as Decision Stump (DS), J48 pruned tree, LMT, Random Forest (RF), SMO, Logitboost, IBK, MultiObjectiveEvolutionary fuzzy (MEF), Multilayer Perceptron (MLP), Logistic regression (LR) comparing the accuracy and also compare values such as Kappa statistic, Mean Absolute Error (MAE), Root Mean Square Error (RMSE). Here 10 fold cross validation method is used for training, validation and testing purpose. Table II represents information about accuracy and Kappa statistic, Mean Absolute Error (MAE), Root Mean Square Error (RMSE) of the different classifier algorithms without selecting features. Table III represents information about accuracy and Kappa statistic, Mean Absolute Error (MAE), Root Mean Square Error (RMSE) of the different classifier algorithms based on attribute selection method for selecting features. Comparison of accuracy, Kappa statistic, Mean Absolute Error, Root Mean square Error without feature selection (WFS) and Feature Selection (FS) of different classification algorithm is shown at fig. 2, fig. 3, fig. 4, fig.5 respectively.

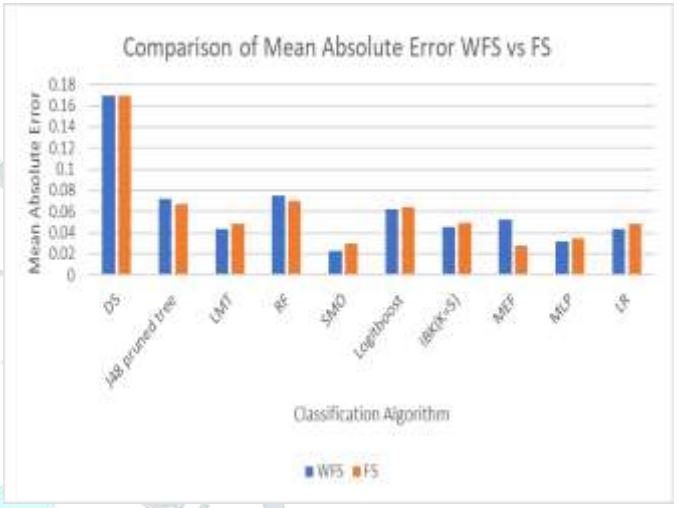

Fig. 4. Comparison of Mean Absolute Error based on WFS and FS.

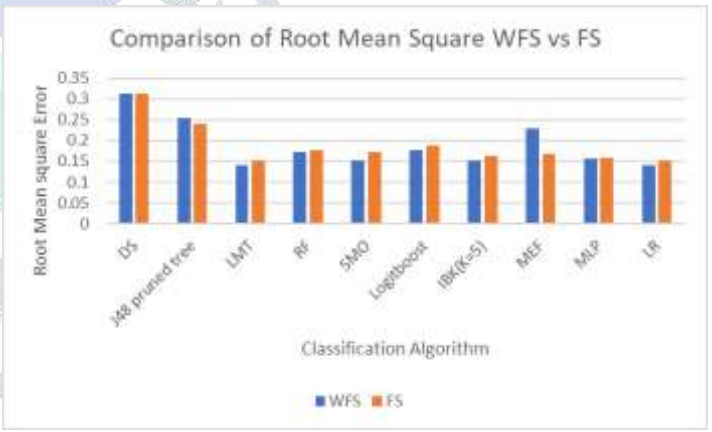

Fig. 5. Comparison of Root Mean square Error based on WFS and FS.

V. Conclusions and Future Work

In this paper at first 10 machine learning algorithms are used on without feature extraction. The value in Table III compares accuracy, kappa statistic, mean absolute error, Root squared of different algorithm based on without feature extraction. Here the accuracy is 97.71% for SMO classification algorithm

performs better also kappa statistics is high and Mean Absolute Error, Root Mean Square Error is also minimum compare to any other algorithm. The value in Table III compares accuracy, kappa statistic, Mean Absolute Error, Root Mean Square Error of different classification algorithm based on CfsSubsetEval attribute selection method. Here the accuracy is 97.53 % for LMT, IBK(K=5), Logistic Regression classification algorithm. Kappa statistics is also high for IBK(K=5) classification algorithm and Mean Absolute Error is minimum for MultiObjective Evolutionary fuzzy (MEF) classification algorithm. Root Mean Square Error is minimum for LMT and Logical Regression Classification algorithm. J48 pruned tree, LMT, MultiObjective Evolutionary fuzzy (MEF), LR classification algorithms perform better on feature extracted attributes. For more accurate result need large dataset. It is concluded that feature analysis and machine learning algorithms play a vital role in determine early diagnosis of breast cancer. For future study different feature selection methods and newer algorithms can be applied to get better results.

References

[1] "Breast cancer", World Health Organization, 2018. [Online]. Available: http://www.who.int/cancer/prevention/diagnosis- screening/breast-cancer/en/. [Accessed: 24- Sep- 2018].

[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. December 2017, pp. 127– 134, 2018.

[3] S. Y. Loke and A. S. G. Lee, "The future of blood-based biomarkers for the early detection of breast cancer," Eur. J. Cancer, vol. 92, pp. 54–68, 2018.

[4] Crisóstomo J, et al. Hyperresistinemia and metabolic dysregulation: the close crosstalk in obese breast cancer. Endocrine. 2016;53(2):433-42.

[5] W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, "Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis," Designs, vol. 2, no. 2, p. 13, 2018.

[6] I. Capasso et al., "Homeostasis model assessment to detect insulin resistance and identify patients at high risk of breast cancer development: National Cancer Institute of Naples experience," J. Exp. Clin. Cancer Res., vol. 32, no. 1, p. 1, 2013.

[7] P. Suryachandra, and P. V. S. Reddy, "Comparison of machine learning algorithms for breast cancer." IEEE International Conference on Inventive Computation Technologies (ICICT), pp. 1- 6, 2016.

[8] Karthik S., Srinivasa Perumal R., Chandra Mouli P.V.S.S.R. (2018) Breast Cancer Classification Using Deep Neural Networks. In: Margret Anouncia S., Wiil U. (eds) Knowledge Computing and Its Applications. Springer, Singapore. https://doi.org/10.1007/978-981-10-6680-1_12.

[9] M. Patrício, J. Pereira, J. Crisóstomo, P. Matafome, M. Gomes, R. Seiça, and F. Caramelo, "Using Resistin, glucose, age and BMI to predict the presence of breast cancer," BMC cancer, vol. 18, no. 1, pp. 29, 2018.

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