RESEARCH ARTICLE OPEN ACCESS

# Comparison of Machine Learning Models for Detecting Breast Cancer Using WEKA Software

Mohit Pant[1], Rashmi Pant[2]

[1] Department of Artificial Intelligence and Machine Learning, Mahakal Institute of Technology, Ujjain, M.P.

[2] Department of Artificial Intelligence and Data Science, Mahakal Institute of Technology, Ujjain, M.P.

***Corresponding author email:** ermohitpant@gmail.com, rashmi.dhanotia@gmail.com

## Abstract

Breast cancer is a major health concern, and early diagnosis is crucial for treatment. In this paper we present a machine learning-based approach for classifying breast cancer using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. CfsSubsetEval with GreedyStepwise is used to select the most relevant features and reduce redundancy. The selected features are evaluated using Logistic Regression, Multilayer Perceptron (MLP) classifiers and Sequential Minimal Optimization (SMO). The models are compared using accuracy, precision, sensitivity, specificity, F1-score, and ROC-AUC. The goal of this study is to identify an efficient classifier and a compact feature subset for reliable breast cancer classification.

Keywords: Breast cancer, Multilayer Perceptron, Regression, Sequential Minimal Optimization

## I. Introduction

Breast cancer is one of the most commonly diagnosed cancers among women worldwide and is a major public-health challenge. Early and accurate diagnosis is essential for the success of clinical decisions and patient outcomes. The increasing availability of computational and biomedical data has led to the development of machine learning-based decision-support strategies to help breast-tumor diagnosis. Machine learning algorithms can analyze numerical diagnostic characteristics to learn patterns associated with benign and malignant tumors. The Wisconsin Diagnostic Breast Cancer (WDBC) dataset is one of the most widely used benchmark datasets for such algorithms. The dataset was developed from digitized fine-needle aspiration images and contains quantitative properties of cell nuclei [1-2].

There have been several machine learning algorithms investigated for breast-cancer classification - Support Vector Machines, Logistic Regression, Decision Trees, Random Forest, k-Nearest Neighbors, Naïve Bayes and Multilayer Perceptron. Similar studies have shown that classification performance can be affected by the different learning techniques [3-4]. However, using all the available features does not necessarily bring the best classification model. The more redundant or less informative features may make the classification process more complicated and may affect the generalization. Therefore, feature selection methods have been studied to locate informative subsets of variables [5-6]. Feature Selection is very important in medical machine learning as it is possible to find a smaller set of relevant variables and this can be advantageous for model interpretation. The previous studies have investigated feature selection using entropy-based feature selection, correlation analysis, principal component analysis and recursive feature elimination [5]–[7]. In this paper, we use Correlation-based Feature Selection (CfsSubsetEval) and GreedyStepwise search strategy to find a small subset of relevant features. CfsSubsetEval evaluates feature subsets by considering the relationship between selected attributes and the target class and thus reducing the redundancy among the attributes. GreedyStepwise searches for a suitable subset of attributes based on the requirement of evaluation [8-9].

## II. Results and Discussion

In this study, we have explored machine learning for breast cancer classification using the Breast Cancer Wisconsin Diagnostic (WDBC) dataset. We have considered data-handling like Logistic Regression, MLP, and SMO in addition to GreedyStepwise feature selection. We observe that feature selection can reduce redundant and less informative features while still delivering good classification performance. Before feature selection, SMO achieved the highest accuracy of 98.25% and MLP and Logistic Regression achieved 95.91% and 90.64%, respectively. After CfsSubsetEval and GreedyStepwise, Logistic Regression showed a big improvement and achieved 98.24% accuracy, 98.60% precision, 96.60% sensitivity, 99.20% specificity, 97.60% F1-score and 0.998 ROC-AUC. MLP and SMO also showed good classification performance with 97.49% and 97.24%, respectively.

We found that correlation-based feature selection and GreedyStepwise search can be effective in finding the relevant and non-redundant features for breast cancer diagnosis. In particular, our improved performance for Logistic Regression shows that a relatively simple and interpretable classifier can get competitive results if the appropriate feature subset is used. Hence, our proposed CfsSubsetEval–GreedyStepwise feature selection framework and Logistic Regression can be considered a promising approach for developing a good breast cancer classification system. Table 2, shows the performance of Logistic Regression, MLP, and SMO before and after using CfsSubsetEval with GreedyStepwise feature selection

| **Parameters** | Values without employed CfsSubsetEval and Greedy Stepwise | | | Values after employed CfsSubsetEval and Greedy Stepwise | | |
|---|---|---|---|---|---|---|
| | **Logistic Regression** | **MLP** | **SMO** | **Logistic Regression** | **MLP** | **SMO** |
| **Correctly Classified Instances** | 155 (90.6433 %) | 164 95.9064 % | 168 98.2456 % | 167 97.6608 | 165 97.4874 % | 166 97.2362 % |
| **Incorrectly Classified Instances** | 16 (9.3567 %) | 7 4.0936 % | 3 1.7544 % | 4 2.3392% | 6 2.5126 % | 5 2.7638 % |
| **Kappa statistic** | 0.8002 | 0.9106 | 0.9622 | 0.9622 | 0.9458 | 0.9403 |
| **Mean absolute error** | 0.0954 | 0.0401 | 0.0175 | 0.0319 | 0.0323 | 0.0276 |
| **Root mean squared error** | 0.3065 | 0.1854 | 0.1325 | 0.1261 | 0.13 | 0.1662 |
| **Relative absolute error %** | 20.3932 | 8.5686 | 3.7492 | 6.8181 | 6.9032 | 5.9141 |
| **Root relative squared error** | 63.3435 | 38.3082 | 27.3699 | 26.0978 | 26.8967 | 34.3983 |

| | | | | | | |
|---|---|---|---|---|---|---|
| % | | | | | | |
| Total Number of Instances | 171 | 171 | 171 | 171 | 171 | 171 |

| Performance Metrics | Value before CfsSubsetEval and Greedy Stepwise employed | | | Value after employed CfsSubsetEval and Greedy Stepwise | | |
|---|---|---|---|---|---|---|
| | Logistic Regression | MLP | SMO | Logistic Regression | MLP | SMO |
| Accuracy (%) | 90.64 | 95.91 | 98.25 | 98.24 | 97.49 | 97.24 |
| Precision (%) | 87.50 | 100 | 100 | 98.60 | 94.59 | 93.60 |
| Sensitivity (%) | 87.50 | 89.10 | 95.31 | 96.60 | 94.59 | 99.20 |
| Specificity (%) | 92.52 | 100 | 100 | 99.20 | 99.20 | 99.20 |
| F1-Score (%) | 87.50 | 94.20 | 97.60 | 97.60 | 94.59 | 96.20 |
| ROC-AUC | 0.949 | 0.997 | 0.977 | 0.998 | 0.987 | 0.966 |

Table: 1 Comparative performance of Logistic Regression, Multilayer Perceptron (MLP), and Sequential Minimal Optimization (SMO) before and after applying selection attributes.

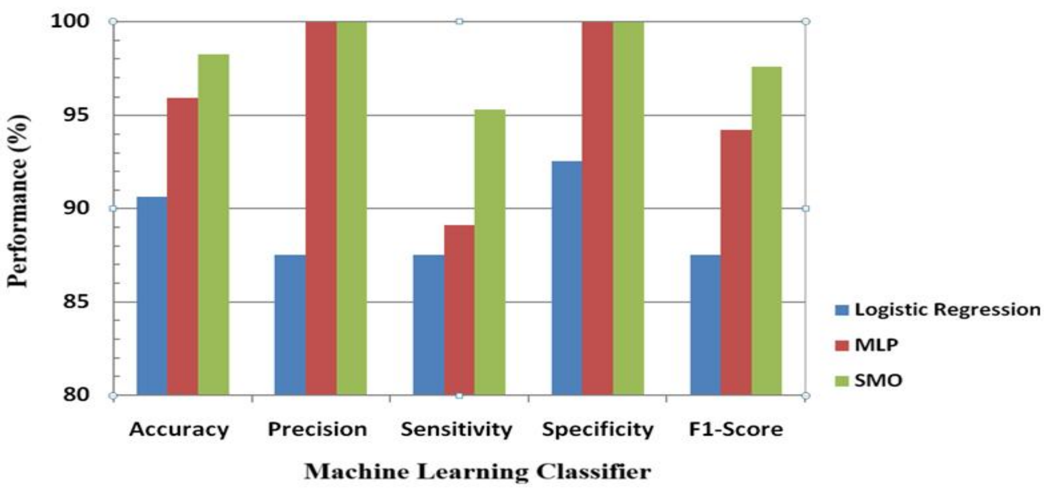


**Figure 1.1 Performance of ML algorithm without using CfsSubsetEval and Greedy Stepwise**

Table 2 Performance metrics of ML algorithms before and after applying selection attributes

The performance is evaluated with accuracy, precision, sensitivity, specificity, F1-score and ROC-AUC. Before feature selection, SMO has the best accuracy of 98.25%, followed by MLP at 95.91% and Logistic Regression at 90.64%. SMO had a high sensitivity of 95.31%, specificity of 100%, F1-score of 97.60% and ROC-AUC of 0.977. MLP has the highest ROC-AUC of 0.997 before feature selection. In the case of CfsSubsetEval with GreedyStepwise, Logistic Regression's accuracy is increased from 90.64% to 98.24% and its precision, sensitivity, specificity, F1-score and ROC-AUC have also increased drastically. However, MLP achieved 97.49% accuracy while SMO achieved 97.24% after feature selection. As shown in the results, feature selection improved the performance of Logistic Regression while maintaining the competitive performance of MLP and SMO.

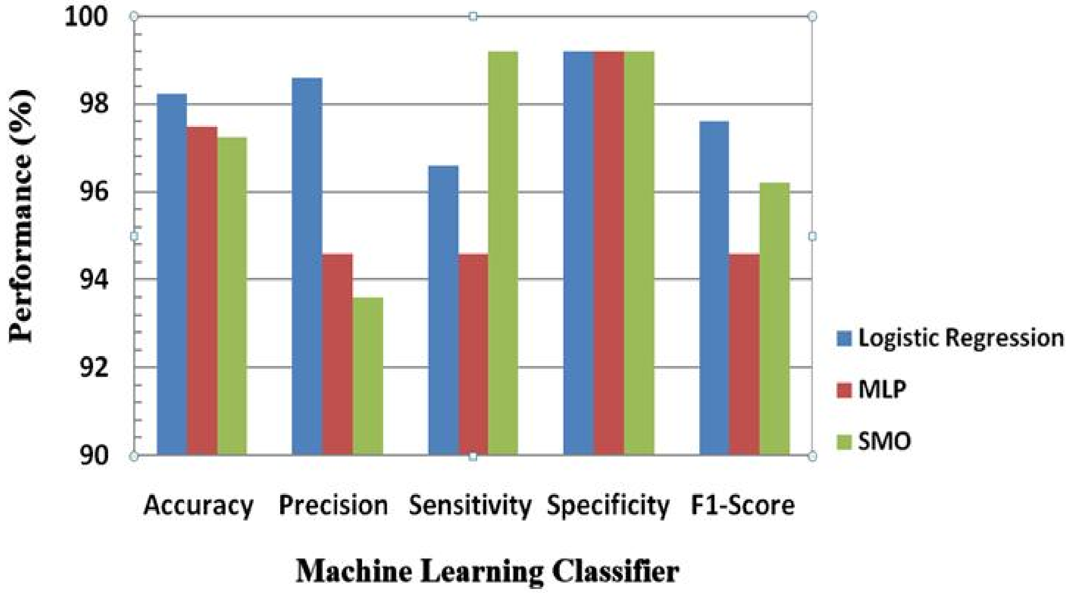


**Figure 1.2 Performance of ML algorithm using CfsSubsetEval and Greedy Stepwise**

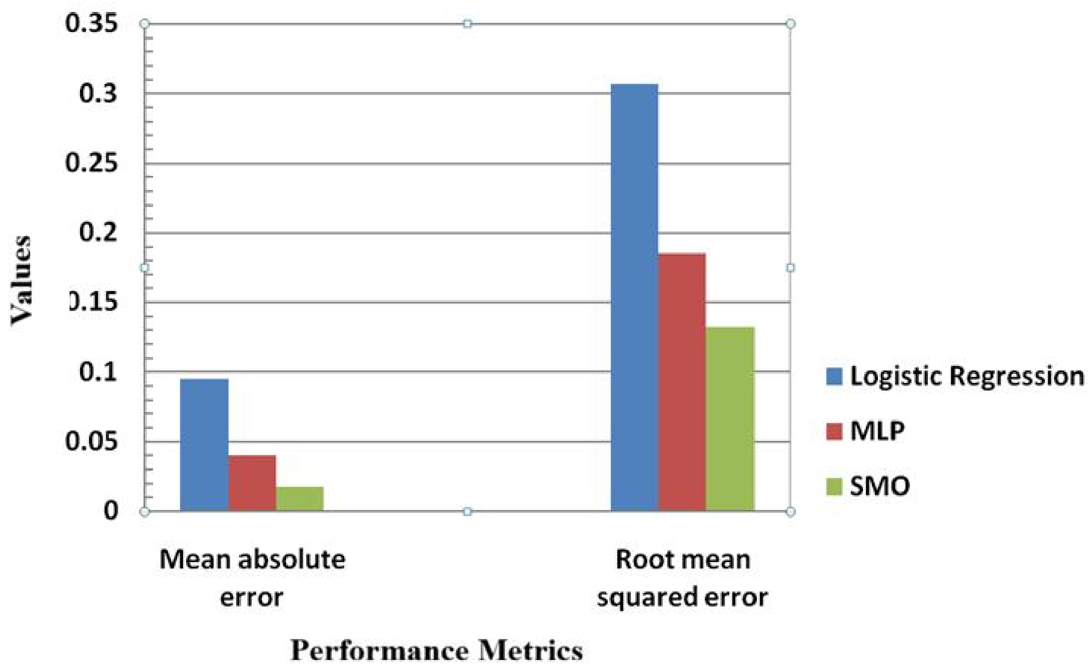


**Fig. 1.3 Performance metric of ML algorithm without using CfsSubsetEval and Greedy Stepwise**

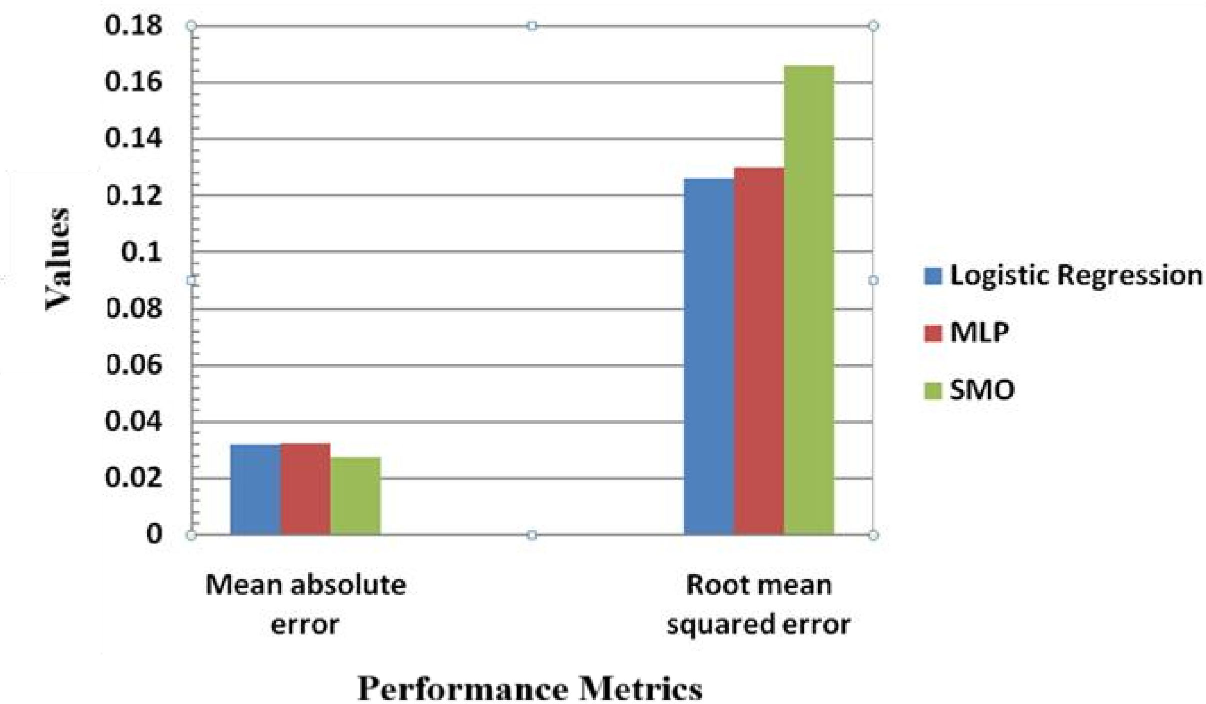


**Fig. 1.4 Performance metric of ML algorithm without using CfsSubsetEval and Greedy Stepwise**

# Conclusion

This paper has investigated the use of machine learning for breast cancer classification using the Breast Cancer Wisconsin Diagnostic (WDBC) dataset. Logistic regression, multilayer perceptron (MLP), and SMO were applied before and after feature selection with CfsSubsetEval with GreedyStepwise. The results show that feature selection can reduce redundant and less informative attributes and still deliver good classification performance.

Before the feature selection, SMO had the highest accuracy of 98.25% and MLP and Logistic Regression had 95.91% and 90.64%, respectively. Logistic Regression was improved with CfsSubsetEval and GreedyStepwise and achieved 98.24% accuracy, 98.60% precision, 96.60% sensitivity, 99.20% specificity, 97.60% F1-score, and 0.998 ROC-AUC. MLP and SMO had very good classification performance and had 97.49% and 97.24% accuracy, respectively.

The correlation-based feature selection combined with GreedyStepwise search can identify the relevant and non-redundant features for breast cancer diagnosis. The better performance of Logistic Regression shows that a simple and interpretable classifier can accomplish competitive results if the right feature subset is used. Therefore, our proposed CfsSubsetEval–GreedyStepwise feature selection framework combined with Logistic Regression can be considered a promising approach for the development of an efficient breast cancer classification system.

# References

**[1]** W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Breast Cancer Wisconsin (Diagnostic)," UCI Machine Learning Repository, 1995.

**[2]** W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates," *Cancer Letters*, vol. 77, no. 2–3, pp. 163–171, 1994.

**[3]** I. Ozcan, H. Aydin, and A. Cetinkaya, "Comparison of classification success rates of different machine learning algorithms in the diagnosis of breast cancer," *Asian Pacific Journal of Cancer Prevention*, vol. 23, no. 10, pp. 3287–3297, 2022.

**[4]** A. F. M. Agarap, "On breast cancer detection: An application of machine learning algorithms on the Wisconsin diagnostic dataset," in *Proc. 2nd Int. Conf. Machine Learning and Soft Computing*, 2018, pp. 5–9.

**[5]** M.-L. Huang, Y.-H. Hung, and W.-Y. Chen, "Neural network classifier with entropy based feature selection on breast cancer diagnosis," *Journal of Medical Systems*, vol. 34, no. 5, pp. 865–873, 2010.

[**6**] Mert, A., Kılıç, N., Bilgili, E., & Akan, A. (2015). Breast cancer detection with reduced feature set. Computational and mathematical methods in medicine, 2015(1), 265138.

**[7]** Ibrahim, S., Nazir, S., & Velastin, S. A. (2021). Feature selection using correlation analysis and principal component analysis for accurate breast cancer diagnosis. Journal of imaging, 7(11), 225.

**[8]** I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Morgan Kaufmann, 2016.

[**9**] Karatza, P., Dalakleidi, K., Athanasiou, M., & Nikita, K. S. (2021, November). Interpretability methods of machine learning algorithms with applications in breast cancer diagnosis. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 2310-2313). IEEE.