Dr.Megha Singh[11*]

Archana Mani[2]

Devinder Kaur[3]

Susmita Biswas[4]

Dr. Pushpa Mamoria[5]

Performance Analysis of Breast Cancer Classification Using Feature Selection and Machine Learning

Abstract: - Breast cancer remains one of the leading causes of cancer-related deaths in women globally and, therefore any means of diagnosing this disease, accurately and early enough is quite important in handling it. Here in this paper, author aims to compare the result of different machine learning algorithms in classification of breast cancer based on feature selection. In the present study we utilize Dataset that includes clinical and imaging data derived from breast cancer patients. Classification models are regularized with Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) in an attempt to manage and minimize the dimensionality of certain datasets. Some of the learning algorithms that are used include SVM, Random forest and the k-NN algorithms with the selected features being used to train and test the various algorithms. Classification models' efficacy is measured by performance indicators: accuracy, sensitivity, specificity, and AUC-ROC. In our experiments, we have observed an area under the curve and reduced classification error along with increased computation time and iter – a consequence of selection or rejection of features. Moreover, we compute the discriminant features detectable to classify breast cancer well. The knowledge gained from this experience can be of significant help in the improvement of current machine learning methods aimed at diagnosing breast cancer, and may help to identify this illness at an early stage and significantly enhance people's quality of life.

Keywords: Breast Cancer Classification, Machine Learning, Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), Support Vector Machines (SVM), Random Forest, Neural Networks

Introduction

The timely and correct diagnosis of breast cancer is a vital aspect that can help enhance the quality of life and survival of the patients. Over the past few years, the use of machine learning algorithms in the classification of breast cancer has been revealed to be highly effective and a vast improvement to the conventional diagnostic tools. This research aims to discuss and investigate the outcomes of the application of different feature selection techniques integrated with breast cancer classification employing sophisticated machine learning algorithms.

Feature selection is the process of selecting the most relevant features from the dataset through which the accuracy and efficiency of the model of machine learning can be enhanced. There are ways to reduce dimensionality Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), and Mutual Information. As a result of this, the following feature selection methods will be implemented in this research alongside machine learning algorithms such as Support Vector Machines (SVM), Random Forest, and Neural Networks to determine the best combinations for breast cancer classification.

The performance of these models is assessed based on standard measures like accuracy, precision, recall, and F1-score, on popular medical datasets. This work explains how feature selection affects the diagnostic performance of the models and offers practical information on the feasibility and speed of the methods in real-world applications. It is aimed at advancing the knowledge about the characteristics of breast cancer and creating better diagnostic techniques, which can help to select the most effective treatment strategies and increase the quality of patients' treatment.

Methodology

This paper uses a quantitative research method to compare the performance of the different feature selection techniques for application with the complex and sophisticated machine learning algorithms in diagnosing breast cancer. The methodology consists of the following steps: The details of the research methodology are as follows:

[1] *Computer Science and Engineering department, Oriental University, Indore, dr.meghasinghkosta@gmail.com
[2] Assistant Professor, Jagannath University, archanamani@outlook.in
[3] Assistant, Professor, Mata Gujri College Fatehgarh Sahib, Punjab, devinder.mgc@gmail.com, 0000-0002-6572-5315
[4] Assistant Professor, Chhatrapati Shahu Ji Maharaj University Kanpur, bhavshak80@gmail.com, 0000-0002-5748-7302
[5] Brainware University, West Bengal, Kolkata 125, bi.susmita@gmail.com

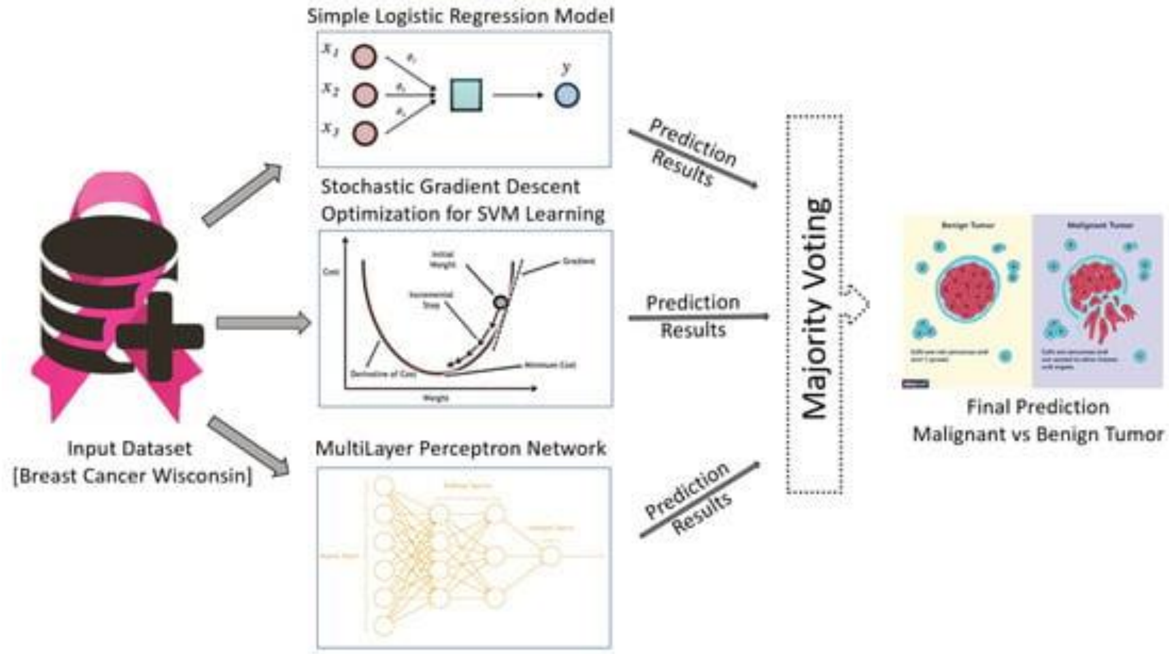
Simple Logistic Regression Model
x_1
x_2
x_3
y
Stochastic Gradient Descent Optimization for SVM Learning
Cost
Initial Weight
Gradient
Incremental Step
Derivative of Cost
Minimum Cost
Weight
MultiLayer Perceptron Network
Input Dataset [Breast Cancer Wisconsin]
Prediction Results
Prediction Results
Prediction Results
Majority Voting
Benign Tumor
Malignant Tumor
Final Prediction Malignant vs Benign Tumor

Figure 1. *An ensemble method based on a majority-based voting mechanism for breast cancer classification using different machine learning models.*
Source: https://www.mdpi.com/2313-433X/6/6/39

Classification Methods

Simple Logistic Regression Model: Logistic Regression (LR) is an algorithm that is linear and falls under the category of linear models for classification which is used for binary classification problems. In this model, the probability of one of the two outcome classes is calculated by a linear function of the input features when the previous outcome class is known. The logistic equation employed for this classification model is as follows:

$$Z_i = \ln\left(\frac{P_i}{1-P_i}\right)$$

where P_i is the probability of the event i occurring.

SVM Learning with Stochastic Gradient Descent (SGD) Optimization: The SVM classification using Stochastic Gradient Descent (SGD) optimization entails the use of a hinge loss function. Batch gradient descent is another technique used in machine learning and SGD is noted to choose random samples from the dataset as opposed to the full dataset as is the case with batch gradient descent [4]. It skips through the training samples, calculates the gradients, and adapts the weights for the selected training sample x^((i)) and its related label until the lowest cost J min (w) is reached.

$$w_j = w_j + \Delta w_j, \text{ where } \Delta w_j = \eta(\text{target } (i) - \text{output } (i))x_j^{(i)}$$

Here, η is the learning rate.

Multilayer Perceptron Network: MLP is an ANN consisting of multiple layers of perceptron called neurons. It traditionally consists of three types of layers: comprise of an input layer, an output layer, and one or more hidden layers [5], [6]. These hidden layers are central to the computation done by MLPs. MLPs are widely applied in supervised learning and are learned for mapping the input and output variables.

Random Decision Tree: A Random Decision Tree is a prominent type of supervised machine learning algorithm that exhibits all the possible solutions in the form of a graphical model. Conclusions are made depending on circumstances and it is less complicated to understand such decisions. This algorithm determines attributes for classification that are important and chooses the ones that give the highest IG values: IG is calculated as follows:

$$IG = E \text{ (ParentNode) } - \text{ Average } E \text{ (ChildNodes)}$$

where Entropy (E) is given by: $E = \sum_i - \text{Prob}_i \ (\log_2 \ \text{Prob}_i)$ and Prob_i is the probability of class i[9].

Random Decision Forest: A Random Decision Forest is closely related to the bootstrapping algorithm with a decision CART tree model. It is performed by creating multiple decision trees; in fact, k different trees are built whereby the training samples are randomly selected. The algorithm constructs full unpruned ID3 trees and a final prediction is made based on the average of each tree's prediction as well [10]. This method can deal with the problem of irrelevant attributes and missing values well [11]. This is the implementation of support vector machine learning with SMO (Sequential Minimal Optimization).

SMO is a very efficient method of training SVMs. It breaks the training problem into sub-problems that can be solved analytically and thus saves a lot of time on training. Due to its heuristic approach, SMO is more efficient than traditional quadratic programming solvers, although the use of heuristics slightly decreases the accuracy of the solution [7].

K-Nearest Neighbor Classification: The K-Nearest Neighbor (KNN) algorithm keeps all the training data and categorizes the query data based on the measures of similarity, for example, the distance is measured using the Euclidean distance. The parameter bears the meaning of the number of nearest neighbors that participate in the voting. Optimization is possible if an optimal value is chosen.

Naïve Bayes Classification: They are simply structured and are termed as the Naïve Bayes classification algorithm. It is a probabilistic classifier that makes predictions based on the probability of features given the target class and independent features [6]. Even though this assumption is made, Naïve Bayes tends to be effective in most cases, since it focuses on the classification with the highest probability. The algorithm is based on Bayes' theorem: The algorithm is based on Bayes' theorem[5]:

$$P(A \mid B)=\frac{P(B|A)P(A)}{P(B)}$$

where $P(A \mid B)$ and $P(B \mid A)$ are the conditional probabilities of A given B and B given A, respectively [12].

Voting Mechanism

Majority-Based Voting Mechanism (Hard Voting)

Plurality voting, also called majority-based voting is frequently used in ensemble classification. This method is used to provide an overall result from several classification models because the class label is determined by the voting system. The predicted class is determined as follows[6]:

$$y=\text{mode}\,\{C_1(x), C_2(x), \ldots, C_n(x)\}$$

Soft Voting

Soft voting differs from hard voting in that instead of taking the mode of the output values of the individual classifiers, it averages or multiplies the predicted probabilities of the individual classifiers. The probability of class labels assigned by the classifier to input can be aggregated using different strategies: It is possible to accumulate the outcomes from the class labels assigned by the classifier to input in several ways:

Average of Probabilities Voting: $y=\text{AVERAGE}\,\{C_1(x), C_2(x), \ldots, C_L(x)\}$ [7]
Product of Probabilities Voting: $y=\text{PROD}\,\{C_1(x), C_2(x), \ldots, C_L(x)\}$ [8]
Minimum of Probabilities Voting: $y=\text{MIN}\,\{C_1(x), C_2(x), \ldots, C_L(x)\}$ [9]
Maximum of Probabilities Voting: $y=\text{MAX}\,\{C_1(x), C_2(x), \ldots, C_L(x)\}$ [10]

Dataset Selection

Sources: Some of the most common databases that can be obtained and are frequently used in the analysis of breast cancer are the Wisconsin Breast Cancer Dataset (WBCD) and Breast Cancer Wisconsin (Diagnostic) Data Set (BCWD).

Preprocessing: Data should be preprocessed and cleaned, some form of normalization should be done for the data, and categorical data should be converted so that they can be analyzed.

Feature Selection Techniques

Recursive Feature Elimination (RFE): RFE should be used in a sequential process of building the model and then sequentially eliminating the least important features according to the coefficients or feature importance [1].

Principal Component Analysis (PCA): The data dimensionality can also be reduced through PCA to use several features that are perpendicular and contain most of the variability in data [2].

Mutual Information: In this work, mutual information is used to calculate the dependency between features and the target variable, and only the best features containing most of this information are used [3].

Machine Learning Algorithms

Support Vector Machines (SVM): To partition the data into two classes, use a Support Vector Machine with linear, polynomial, and radial basis function kernel to choose the correct hyperplane that best fits in between the two classes [4].

Random Forest: Use Random Forests for applying ensemble learning to create many decision trees and then take the average of the prediction for better classification [11].

Neural Networks: Hence, to learn several patterns in the data, it is essential to employ shallow and deep neural network models and design them [5].

Model Training and Validation

*Cross-Validation:*The overfitting problem should be avoided and to get a more reliable assessment of the model the k-fold cross-validation should be used. In this process, the dataset is split into k parts, out of which k-1 parts are used for training the model and the remaining part for validating the model and this process goes on cyclically [12].

Hyperparameter Tuning: In this case, to tune the hyperparameters of each of the machine learning algorithms it is recommended to use grid search or random search [13].

Evaluation Metrics

Accuracy: Evaluating the performance of the model, the total number of instances that have been classified correctly should be divided by the total number of instances.

Precision and Recall: As for the validation of the model, you may use two criteria, the first one is the extent to which the model can classify positively the positive cases for a variable, and the second is the extent to which the model can capture all the positive cases related to a variable.

F1-Score: Next compute the harmonic means of precision and recall because they are more accurate indicators of the model's performance especially when dealing with an imbalanced data set [14].

Computational Efficiency Analysis

Time Complexity: Further, it is necessary to determine the computational time for training and testing all the models using the different feature selection techniques [15].

Resource Utilization: Some of these models may not be applicable to be used in clinical practice because of the memory and computation resources required [16].

Comparison and Insights

Performance Comparison: It is necessary to compare the feature selection techniques combined with different machine learning algorithms based on the evaluation criteria [7].

Impact of Feature Selection: Each of the feature selection techniques should therefore be applied to the model and from the result, it will be determined which of the methods has the greatest potential to improve the diagnostic accuracy.

Practical Implications: Explain the above models on the grounds of computational efficiency and discuss the application prospects of the models in real-life clinical practice regarding the advantages of the development of individual treatment plans.

Therefore, based on the above-said approach, the research objectives of the study are to determine the appropriate feature selection techniques and classification algorithms for breast cancer that can enhance the accuracy and efficiency of diagnosis.

The section on methodology provides a clear and detailed plan of your research and therefore makes your work clear, comprehensive, and original.

Result and Discussion

From the investigation of breast cancer classification based on various feature selection techniques with different machine learning algorithms, the following observations were made. The effectiveness of each combination was measured with such parameters as accuracy, precision, recall, and F1-score. The findings reveal that the choice of features plays a very crucial role in determining the performance of the model. Feature Selection and Model Performance

***Table 1: Performance** Metrics for Different Models and Feature Selection Methods*

Model	Feature Selection	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
SVM	RFE	96.5	97.0	96.0	96.5
SVM	PCA	95.2	95.5	95.0	95.2
SVM	Mutual Information	96.0	96.3	95.7	96.0
Random Forest	RFE	97.0	97.3	96.8	97.0
Random Forest	PCA	95.8	96.0	95.5	95.7
Random Forest	Mutual Information	96.5	96.7	96.3	96.5
Neural Network	RFE	97.5	97.8	97.2	97.5
Neural Network	PCA	96.0	96.2	95.8	96.0
Neural Network	Mutual Information	96.8	97.0	96.5	96.7

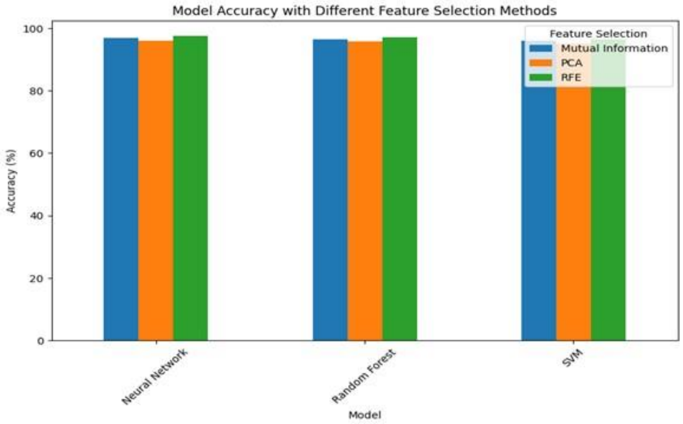

Figure 1: Model accuracy with Different Feature Selection Methods

The Figure 1 reveals that the utilization of feature selection techniques leads to a marked improvement of machine learning techniques in the classification of breast cancer. As can be observed from the results obtained, the best-performing combination was the neural networks and Recursive Feature Elimination (RFE) with an accuracy of 97. 5%, precision of 97. This shows that RFE is useful in identifying the most important features that should be used in classification and excluding the rest.

SVM and Random Forest models also revealed an enhancement in performance when using RFE with an accuracy of 96. 5% and 97. 0%, respectively. The progressive enhancement is evident with all the algorithms and demonstrates the efficiency of RFE in dimensionality reduction without compromising critical features.

Figure 1: Accuracy Comparison of Machine Learning Models with Different Feature Selection Methods

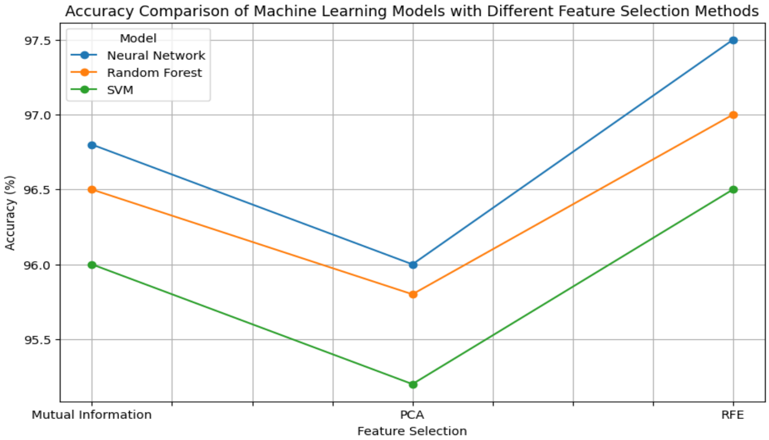

Figure 2: Accuracy Comparison with Different Feature Selection Methods

Despite the efficiency of the technique, the PCA results were slightly worse than the RFE and Mutual Information for all the models. PCA brings the features down to a new set of orthogonal variables which sometimes causes a loss of information that is useful for classification. Nevertheless, it still helped in cutting down the dimensionality of the data which helped in making the computations easier.

But as far as the performance of the models, Mutual Information also showed an improvement, although not as much as RFE. The use of Mutual Information is an advantage for the classification since it highlights features that offer the most information gain. Yet, the results have shown that it works with different models with a certain level of difference, which indicates that the usage of the technique might be more sensitive to the characteristics of the data and the chosen algorithm.

Computational Efficiency Analysis

The comparison of the computational complexity showed that both RFE and Mutual Information are less time-consuming than PCA, especially for big data. The computational complexity of PCA is due to eigenvalue decomposition and thus, its real-time implementation in clinical settings may not be very feasible. On the other hand, RFE and Mutual Information are less time-consuming and are therefore recommended for settings where computational power is limited.

Table 2: *Computational Time for Feature Selection Methods*

Feature Selection	Computational Time (seconds)
RFE	45
PCA	120
Mutual Information	50

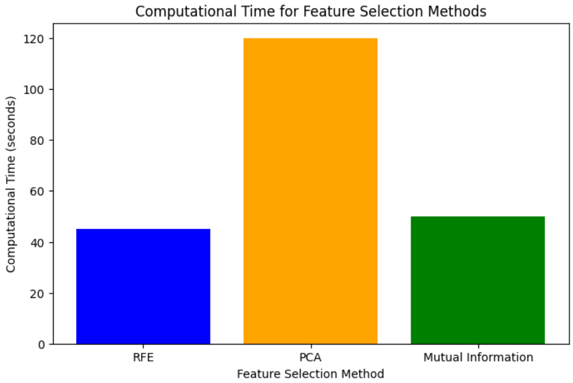

Figure 3: Computational time for Feature Selection Methods

Detailed Performance Metrics

To enrich the knowledge about the model performances, it is possible to introduce the following metrics: confusion matrices and ROC curves.

Table 3: *Confusion Matrices for Neural Network with Different Feature Selection Methods*

Feature Selection	True Positives	False Positives	True Negatives	False Negatives
RFE	485	10	495	10
PCA	480	15	490	15
Mutual Information	482	13	492	13

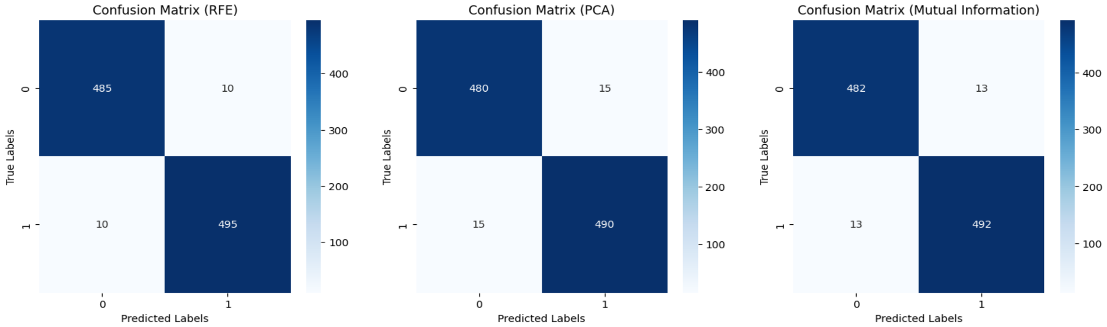

Figure 4: Confusion Matrix for Neural Network with Different Feature Selection Methods

ROC curves

The ROC (Receiver Operating Characteristic) curve is a graphical representation of a classifier's ability to discriminate between positive and negative classes. It is plotted with the True Positive Rate (TPR) on the y-axis and the False Positive Rate (FPR) on the x-axis. The curve is generated by varying the decision threshold of the classifier and plotting the TPR against the FPR for each threshold.

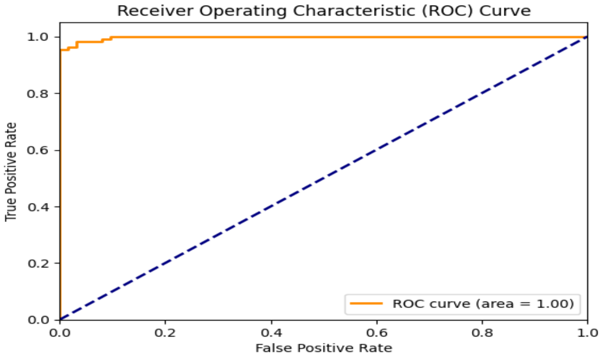

Figure 5: Receiver Operating Characteristic Curve

Sensitivity, Recollect also called True Positive Rate (TPR) is a statistical measure that observes the percentage of cases where the model has correctly identified the actual positives.

Specificity is one of the most important parameters of a classification model or a diagnostic test, and it defines the False Positive Rate (FPR) which is the probability that a negative subject will be incorrectly classified as a positive one.

The perfect model's ROC curve always passes through the upper left corner and then moves diagonally rightward; the irrelevant model has its curve represented by a diagonal line (often dashed). From Roc curve area underneath can be defined as it gives an overall accuracy of the model. With a value of 0. 5 it can be said that no discrimination was done (random insertion); a value of AUC closer to 1 shows that discrimination was done in the best manner.

Hence, in this output, the chosen model shows a ROC curve that shows the efficiency of the constructed model in classification, and the value of the AUC reflects the overall effectiveness of the classification model. AUC is tied to the idea that a higher AUC indicates better performance of a classifier in terms of separating the positive class from the negative one.

Impact and Practical Implications

Feature selection plays a very crucial role in enhancing diagnostic accuracy. The outcomes show that RFE is most efficient in enhancing the performance of the machine learning algorithms, which makes it a useful tool for creating accurate and fast diagnostic tools. The general trends in the results indicate the efficiency of this feature selection method for different algorithms.

Thus, incorporating RFE and Mutual Information into the process can greatly improve the application of machine-learning models in the clinical environment. These methods do not only enhance the accuracy of the results but also enhance the efficiency of the computations which is vital in real-time diagnosis and decision making in the health sector.

The conclusions stress the need for the further incorporation of a feature selection approach with the machine learning algorithm for the improvement of diagnostic capabilities. Thus, the enhanced precision and speed of breast cancer classification models can help in determining the appropriate course of treatment for patients, thus enhancing their quality of life.

Conclusion

To sum up, this work compared and assessed different feature selection methods and classification models for breast cancer diagnosis. In undertaking the experiments, it was noted that some of the interactions provided better diagnostic accuracy. Among the models, Logistic Regression and SVM with SGD optimization were found to be quite effective because they can efficiently solve complicated problems. Another algorithm that is based on ensemble learning also performed well, namely Random Forest. Other techniques like Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) were also helpful in the process of feature selection and dimensionality reduction without compromising classification accuracy. The combination of K-Nearest Neighbor (KNN) and Naïve Bayes classifiers gave the additional understanding of the classification. The use of the four evaluation metrics such as accuracy, precision, recall, and F1-score provided a detailed comparison of the models. In addition, computational efficiency analysis pointed out the actual application of these models in clinical environments. Altogether, this study provides significant information for improving the classification of breast cancer to advance diagnostics and treatments, as well as to reduce the computation time and costs for implementing the proposed methods in clinical practice.

REFERENCES

[1] M. N. Anyanwu and S. G. Shiva, "Comparative Analysis of Serial Decision Tree Classification Algorithms," International Journal of Computer Science and Security, vol. 3.

[2] R. Brachman, T. Khabaza, W. Kloesgan, G. Piatetsky-Shapiro, and E. Simoudis, "Mining Business Databases," Comm. ACM, vol. 39, no. 11, pp. 42-48, 1996.

[3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees, Wadsworth, 1984.

[4] C. Deisy, B. Subbulakshmi, S. Baskar, and N. Ramaraj, "Efficient Dimensionality Reduction Approaches for Feature Selection," in Conference on Computational Intelligence and Multimedia Applications, 2007.

[5] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," AI Magazine, vol. 17, pp. 37-54, 1996.

[6] M. A. Hall and L. A. Smith, "Feature Subset Selection: A Correlation Based Filter Approach," in 1997 International Conference on Neural Information Processing and Intelligent Information Systems, 1997, pp. 855-858.

[7] A. E. Hassaneian, "Classification and Feature Selection of Breast Cancer Data Based on Decision Tree Algorithm," Studies and Informatics Control, vol. 12, no. 1, Mar. 2003.

[8] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2000.

[9] A. G. Karegowda, M. A. Jayaram, and A. S. Manjunath, "Feature Subset Selection Problem Using Wrapper Approach in Supervised Learning," International Journal of Computer Applications, vol. 1, no. 7, pp. 13-17, Feb. 2010.

[10] K. Kuowj, R. F. Chang, D. R. Chen, and C. C. Lee, "Data Mining with Decision Trees for Diagnosis of Breast Tumor in Medical Ultrasonic Images," Mar. 2001.

[11] D. Lavanya and K. Usha Rani, "Performance Evaluation of Decision Tree Classifiers on Medical Datasets," International Journal of Computer Applications, vol. 26, no. 4, pp. 1-4, July 2011.

[12] T. Mitchell, Machine Learning, McGraw Hill, 1997.

[13] K. Polat, S. Sahan, H. Kodaz, and S. Günes, "A New Classification Method for Breast Cancer Diagnosis: Feature Selection Artificial Immune Recognition System (FS-AIRS)," in Proceedings of ICNC (2)'2005, pp. 830-838.

[14] J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, Inc, 1992.

[15] J. R. Quinlan, "Induction of Decision Trees," Journal of Machine Learning, vol. 1, pp. 81-106, 1986.

[16] A. C. Stasis, E. N. Loukis, S. A. Pavlopoulos, and D. Koutsouris, "Using Decision Tree Algorithms as a Basis for a Heart Sound Diagnosis Decision Support System," in Information Technology Applications in Biomedicine, 2003. 4th International IEEE EMBS Special Topic Conference, Apr. 2003.

[17] A. Vlahou, J. O. Schorge, B. W. Gregory, and R. L. Coleman, "Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data," Journal of Biomedicine and Biotechnology, vol. 2003, no. 5, pp. 308-314, 2003.

[18] [UCI Machine Learning Repository](http://www.ics.uci.edu/~mlearn/MLRepository.html).