ISSN 2616-5791 Vol.4, Issue 3: 25-33, DOI: 10.25236/AJMHS.2023.040305

A Comparative Study of Seven Machine Learning Algorithms for Breast Cancer Detection and Diagnosis

Qinyi Ruan[1,a,*]

[1]*Department of Mathematics and Applied Mathematics, Wenzhou University, Wenzhou, Zhejiang, China*
[a]*20211031233@stu.wzu.edu.cn*
[*]*Corresponding author*

Abstract: *This paper presents a comparative analysis of seven distinct machine learning (ML) algorithms, namely Linear Discriminant Analysis, Logistic Regression, K-Nearest Neighbor, Decision Tree Classifier, Random Forest Classifier, Voting Classifier, and Support Vector Machine, in predicting the diagnosis of breast cancer. The study utilized 30 histological tumor features obtained from digital imaging of fine needle aspirates of breast tumor cell masses contained in the dataset, achieving an accuracy of approximately 95% through the application of the aforementioned algorithms. Results show that the LDA and RFC algorithms outperformed the others in terms of accuracy in diagnosing breast cancer. Furthermore, the study suggests that the stability of diagnostic outcomes is better achieved with large-scale data. Finally, the accuracy of the LR algorithm was observed to be less than 85% after conducting Principal Component Analysis (PCA), which was lower than the accuracy achieved without dimensionality reduction.*

Keywords: *Machine Learning ML, Linear Discriminant Analysis, Logistic Regression, K- Nearest Neighbor, Decision Tree Classifier, Random Forest Classifier, Voting Classifier, Support Vector Machine, Breast cancer, Principal component analysis*

1. Introduction

Breast cancer is a highly fatal disease affecting women worldwide and is currently the second leading cause of death in women, second only to lung cancer. There are two types of tumors associated with breast cancer, namely benign (non-cancerous) and malignant (cancerous), each having distinct risks and treatment options. The onset of breast cancer is marked by uncontrolled cell growth and it is crucial to detect it as early as possible to prevent it from spreading further. Hence, it is imperative to identify the type of tumor accurately at an early stage and administer appropriate treatment promptly.

In recent years, machine learning (ML) techniques have been increasingly utilized for the development of prognostic models. In cancer research, these methods have proven to be valuable in identifying distinctive patterns in data sets and predicting whether a cancer is malignant or benign, thereby assisting physicians and patients in making informed decisions. While several models have been used for disease prediction and quantification, such as K-Nearest Neighbour for linear models and Random Forest Classifier for more complete algorithms, researchers continue to strive for more accurate and suitable diagnostic systems that can enable faster and easier tumor detection, leading to earlier treatment and improved survival rates.

In this study, we investigate and compare seven popular ML techniques applied to breast cancer datasets. These techniques include linear discriminant analysis, logistic regression, K-nearest neighbor, decision tree classifier, random forest classifier, and voting classifier. In the remainder of this article, we provide a literature review of related studies in Section 2, followed by the data and methodology of our study in Section 3. In this section, we construct the dataset and describe all the variables used to analyze the seven ML classification methods under study, including the principal component analysis used. Section 4 presents the experimental analysis, where we first compare and analyze the heat maps generated by all factors in the dataset for the diagnosis of malignant and benign tumors, then scientifically compare the accuracy of the seven machine learning models for breast cancer diagnosis using box plots, and finally compare the accuracy of each model after dimensionality reduction. Finally,

ISSN 2616-5791 Vol.4, Issue 3: 25-33, DOI: 10.25236/AJMHS.2023.040305

Section 5 summarizes the findings of this research.

2. Literature Review

This section presents a review of previous studies in which researchers used different machine learning methods for breast cancer diagnosis.

Ahmad et al. conducted a study to compare the performance of decision trees (C4.5), support vector machines (SVM), and artificial neural networks (ANN) for breast cancer diagnosis. They used a dataset from the Iranian breast cancer center and found that SVM had the highest accuracy, followed by ANN and decision tree. [1]

Ojha and Goel used various machine learning algorithms to predict recurrent breast cancer cases using the Wisconsin Prognostic Breast Cancer (WPBC) dataset. Their evaluation showed that SVM and decision trees (C 5.0) were the best predictors with an accuracy of 81%, while fuzzy c-means had the lowest accuracy of 37%. [2]

Delen et al. built a breast cancer survival prediction model by analyzing a large dataset, the Surveillance, Epidemiology, and End Results (SEER) cancer incidence database, using artificial neural networks, decision trees, and logistic regression. [3]

Mandeep Rana et al. conducted a comparative study of different machine learning techniques such as support vector machine (SVM), logistic regression, K-nearest neighbours (KNN) and simple Bayes to predict breast cancer recurrence and diagnose breast cancer using these techniques. The dataset used in this study is from the Wisconsin Breast Cancer Prognosis Dataset (UCI) repository and all 32 variables were used in this work with 95.6% accuracy for breast cancer detection and 68% accuracy for recurrent and non-recurrent breast cancers. [4]

Agarap in 2018 compared the accuracy of six machine learning (ML) algorithms for breast cancer detection: the GRU-SVM, linear regression, multilayer perceptron (MLP), nearest neighbor search (NN), SoftMax regression, and support vector machine (SVM). The results showed that all the proposed ML algorithms performed well in the classification task (accuracy greater than 90% for all tests). The MLP algorithm outperformed the implemented algorithms with a test accuracy of ≈99.04% [5].

In summary, previous studies have shown that machine learning techniques can be effective in breast cancer diagnosis and survival prediction. SVM, decision trees, and artificial neural networks have been frequently used and shown to produce high accuracy in breast cancer diagnosis. These studies provide a foundation for our research and demonstrate the potential of machine learning algorithms for improving breast cancer diagnosis and treatment.

3. Data and methodology

3.1. Dataset

The dataset for this study was derived from kaggle and had 569 data points, 212-malignant and 357-benign. This dataset contains features obtained from digital imaging of fine needle aspirates of breast tumor cell masses, which are described as:

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

ISSN 2616-5791 Vol.4, Issue 3: 25-33, DOI: 10.25236/AJMHS.2023.040305

j) fractal dimension ("coastline approximation" - 1).

With each feature having three information: (1) mean, (2) standard error, and (3) worst. Therefore, having a total of 30 dataset features.

3.2. Machine Learning (ML) Algorithms

Machine learning is a subfield of computer science that focuses on developing algorithms capable of learning from data without being explicitly programmed. It involves the use of statistical models and computational techniques to analyze and learn patterns from large datasets. In this study, we used a two-phase experimental approach, involving a training phase and a testing phase, to develop a machine learning algorithm capable of accurately distinguishing between benign and malignant tumors for clinical diagnosis.

To achieve this, we evaluated seven different common machine learning algorithms, namely Linear Discriminant Analysis, Logistic Regression, K-Nearest Neighbor, Decision Tree Classifier, Random Forest Classifier, Voting Classifier, and Support Vector Machine. Each of these algorithms has its unique strengths and weaknesses in terms of predictive accuracy, computational complexity, and interpretability. Therefore, we conducted a parametric study of these algorithms to identify the most suitable method for breast cancer diagnosis.

In the training phase, we used 80% of the dataset to train the model, while the remaining 20% was used for testing to prevent overfitting. The goal was to develop a model that accurately predicted the type of tumor (benign or malignant) based on the histological tumor features obtained from digital imaging of fine needle aspirates of breast tumor cell masses. The results of our study provide valuable insights into the performance of different machine learning algorithms for breast cancer diagnosis.

3.2.1. Support Vector Machine(SVM)

Developed by Vapnik [6], the support vector machine (SVM) was primarily intended for binary classification. Its main objective is to determine the optimal hyperplane $f(w,x) = w \cdot x + b$ separating two classes in a given dataset havinginput features $x \in R_p$ and labels $y \in \{-1,+1\}$. SVM learns by solving the following constrained optimization problem:

$$\min \frac{1}{p} w^T w + C \sum_{i=1}^{p} \xi_i \tag{1}$$

$$s.t\, y'_i (w \cdot x + b) \geq 1 - \xi_i \tag{2}$$

$$\xi_i \geq 0, i = 1, \ldots, p \tag{3}$$

Where $w^T w$ is the Manhattan norm, ξ is a cost function, and C is the penalty parameter (may be an arbitrary value or a selected value using hyper-parameter tuning). Its corresponding unconstrained optimization problem is the following:

$$\min \frac{1}{p} w^T w + C \sum_{i=1}^{p} \max\left(0, 1 - y'_i (w_i x_i + b)\right) \tag{4}$$

Where $wx + b$ is the predictor function

3.2.2. K-Nearest Neighbor (KNN)

K-nearest neighbor is a nonparametric dispersion algorithm. Nearest neighbors are selected based on the Euclidean distance between x and y vectors according to equation (5). The results of KNN depend on the value of K [7]. A large value of K leads to overlapping classes, while a small value of K increases the number of computations.

ISSN 2616-5791 Vol.4, Issue 3: 25-33, DOI: 10.25236/AJMHS.2023.040305

$$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \tag{5}$$

3.2.3. Voting Classifier (Voting)

The voting classification strategy is useful when a failure in one classifier algorithm can be advantageous for another. This method combines the anticipated results of multiple classifiers, such as J48 decision trees, SVM, and Naïve Bayes. To preprocess the dataset, attribute ranking is performed using a classification algorithm to eliminate features that do not reach a global minimum. The filtered dataset is then used for each classifier individually and in combination to achieve the highest accuracy. The predicted outputs of each classifier are combined, and the most frequent predicted class is selected as the class variable for the test instances [8].

3.2.4. Random Forest Classifier (RFC)

Similar to how a jury arrives at a judicial decision, Random Forest (RF) combines many decision trees into a set to create a forest of trees. The advantage of using RF is that having a single decision tree can provide a simple or a very specific model [9]. Using RF results in greater stability compared to using a single decision tree, as it is not sensitive to noise in the input data set. In cancer detection, one of the main reasons for using RF is its ability to handle minority classes in the data. For instance, a tumor can be classified as benign or malignant, even though the latter category represents only 10% of the input data set. The RF method is based on a recursive approach in which each iteration picks a random sample of size N from the data set and performs a substitution, while another random sample is picked from the predictor without substitution. The resulting data are split, and the non-partitioned data are discarded. These steps are repeated several times, depending on the number of trees needed. Finally, the cases are sorted based on the majority vote in the decision trees [10].

3.2.5. Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) is a statistical method commonly used for feature extraction to reduce the dimensionality of high-dimensional data. It is primarily used for controlled dimensionality reduction, where the goal is to identify the most relevant features that can discriminate between different classes. LDA achieves this by maximizing the interclass variance and minimizing the intraclass variance simultaneously, which leads to the highest degree of separation between classes [11]. LDA employs the eigen decomposition of the covariance matrix to compute the optimal transformation that maximizes the interclass variance. This allows LDA to identify the features that contribute the most to the differences between classes. Unlike principal component analysis (PCA), LDA often yields better results for classification problems. In this study, we created separate LDA training and LDA validation datasets to train and validate our model.

3.2.6. Decision Tree Classifier (DTC)

Decision tree models comprise decision points, strategy points (or event points), and tree structure results, and are commonly used for decision-making purposes. They are usually employed as a decision criterion for maximizing expected performance or minimizing expected cost by plotting the efficacy of various alternatives under distinct circumstances and comparing them.

3.2.7. Logistic Regression (LR)

Linear regression. Despite the algorithm used for the regression task, linear regression was used as the classifier in this study (see Equation 6). For this purpose, the output threshold of Equation 7 was used, that is, the regression value of Equation 7 was taken.

$$h_\theta(x) = \sum_{i=0}^{n} \theta_i x_i + b \tag{6}$$

$$f(h_\theta(x)) = \begin{cases} 1 & h_\theta(x) \geq 0.5 \\ 0 & h_\theta(x) < 0.5 \end{cases} \tag{7}$$

To measure the loss of the model, the mean squared error (MSE) was used (see Eq. 8).

ISSN 2616-5791 Vol.4, Issue 3: 25-33, DOI: 10.25236/AJMHS.2023.040305

$$L(y,\theta,x)=\frac{1}{N}\sum_{i=0}^{N}(y_i-(\theta_i\cdot x_i+b))^2 \tag{8}$$

where y represents the actual class, and $(\theta \cdot x + b)$ represents thepredicted class. This loss is minimized using the SGD algorithm, which learns the parameters θ of Eq. 6. The same method of lossminimization was used for MLP and Softmax Regression. [13]

3.3. Principal Component Analysis (PCA)

Principal component analysis (PCA) is a feature extraction technique that transforms the original dataset into a reduced set of uncorrelated variables known as principal components (PCs). In this work, we apply PCA to neural networks. PCA commonly employs cumulative variance to reduce the dimensionality of the features in the dataset, selecting the number of principal components based on either the dimensionality of the feature values or the proportion of variance accounted for by each principal component.

4. Empirical Analysis

In Figure 1, Among the 569 breast cancer patients, the number of patients diagnosed with malignant tumors was 212, or 37%, while the number of patients diagnosed with benign tumors was 357, almost 1.7 times the number of the former, or 63%. This illustrates that the odds of being diagnosed with benign tumors are greater than malignant tumors in those who have breast cancer.

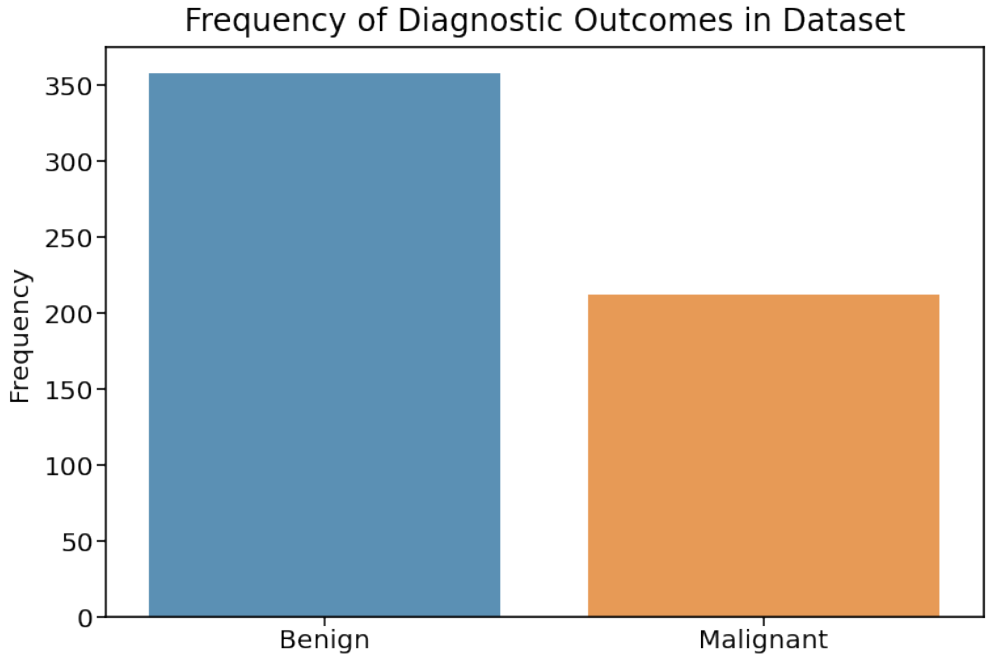

Figure 1: Frequency of Diagnostic Outcomes in Dataset.

Figure 2 shows the heat map generated from the correlation between the factors, and the heat map is symmetrical on both sides with the left diagonal line as the axis of symmetry. From Figure 2, it can be obtained that:

1) The corresponding chart colors between the six factors correlated with radius and perimeter, and the six factors correlated with smoothness and symmetry are dark purple and black, indicating that radius and perimeter are unrelated to smoothness and symmetry.

2) The chart color of standard error of three factors of concavity with radius and perimeter is light purple, which means that there is little correlation between the factors, and the chart color of standard error with smoothness and symmetry is purple and dark purple, which means that the factors are almost unrelated to each other. In particular, the standard error of concavity has very little correlation with

ISSN 2616-5791 Vol.4, Issue 3: 25-33, DOI: 10.25236/AJMHS.2023.040305

each factor of radius, perimeter, smoothness, and symmetry, which is almost unrelated. However, the mean value of concavity and the worst value of concavity are correlated with the remaining factors except for the standard error of radius and perimeter, because the corresponding chart colors appear flesh-colored and orange.

3) The chart colors between the three factors of radius and the three factors of perimeter are white and light flesh color, and even the columns of the same attribute of radius and perimeter have the same color, which indicates a very strong correlation between radius and perimeter.

4) Diagnosis is closely related to the mean value of radius, perimeter and concavity, the worst value of radius, perimeter and concavity, and the mean value of smoothness and symmetry, the standard error of radius and perimeter error, and the worst value of smoothness and symmetry, and almost independent of the standard error of smoothness, concavity, and symmetry.

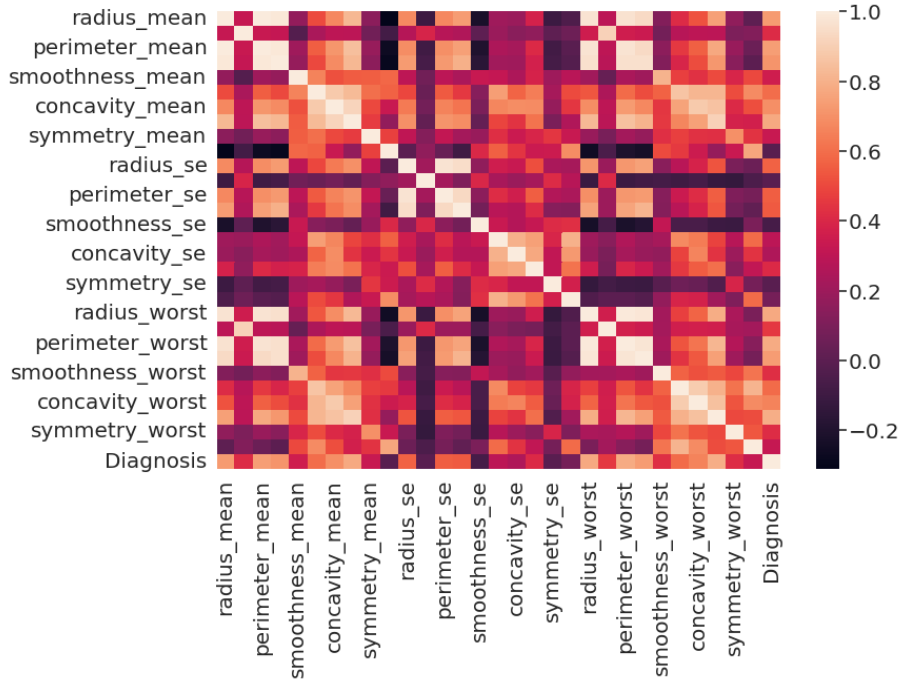

Figure 2: The Heatmap of All Factors.

Figure 3 shows box plots of the accuracy of breast cancer diagnosis using each of the seven machine learning models, and it can be seen that:

1) The median accuracy of the RFC algorithm exceeds 0.950 and is in the highest position compared to the other six algorithms; the maximum value of the LDA algorithm accuracy exceeds 0.9875, which is much higher than the other six algorithms, and it ranks second in the median, which indicate that the accuracy and diagnosis of breast cancer using LDA and RFC are the best.

2) The median positions of DTC, RFC and SVM were all biased toward the lower quartile and all showed outliers, with the accuracy of the outliers being around 0.900, 0.9125 and 0.860, respectively, indicating that the accuracy of the diagnosis may decrease under individual special circumstances.

3) The box-line plots generated by DTC and RFC have the smallest IQR, close to 0.0125, indicating that the two algorithms produce the smallest dispersion, and among the seven algorithms, the stability of the diagnostic results under large-scale data is the best, followed by LDA.

4) The overall position of the box line plots generated by LR and SVM is low, the lowest value of the box line plots of LR is below 0.875, the lowest value of the box line plots of KNN is beyond but close to 0.875, while the lowest values of the box line plots of several other models are almost all at 0.900 and above, so the accuracy of the diagnosis with LR and SVM is low relative to the other five algorithms.

ISSN 2616-5791 Vol.4, Issue 3: 25-33, DOI: 10.25236/AJMHS.2023.040305

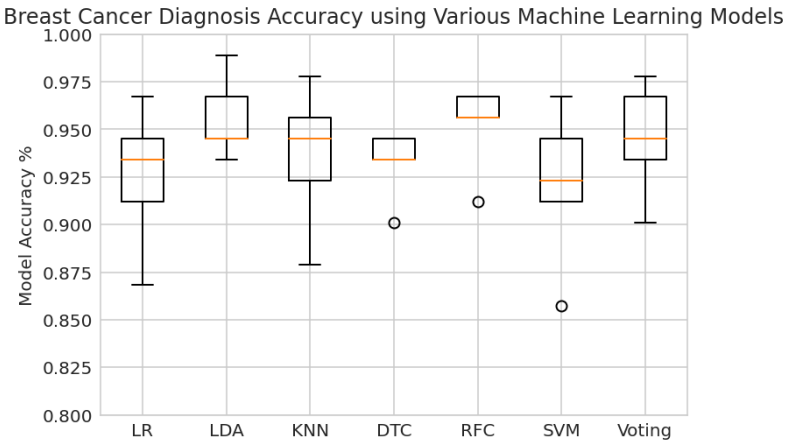

Figure 3: Breast Cancer Diagnosis Accuracy using Various Machine Learning Models.

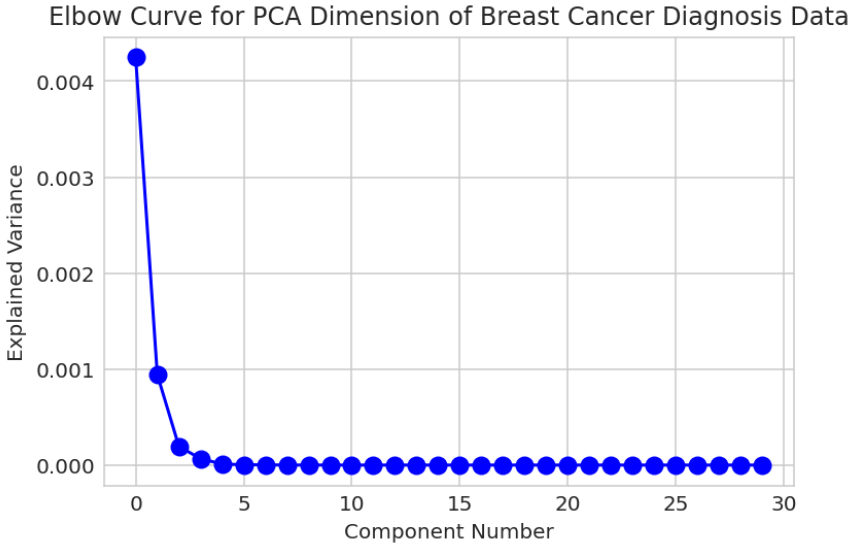

Figure 4: Elbow Curve for PCA Dimension of Breast Cancer Diagnosis Data.

Figure 4 shows the Elbow Curve for PCA Dimension of Breast Cancer Diagnosis Data, where we used the dimensionality reduction method to reduce the 30 features to the principal components (PCA) to maximize the explanation of the differences in the data. As can be seen in Figure 4, almost only the first three variables are valid. Figure 5 shows the Data Visualized After 3-Component PCA.

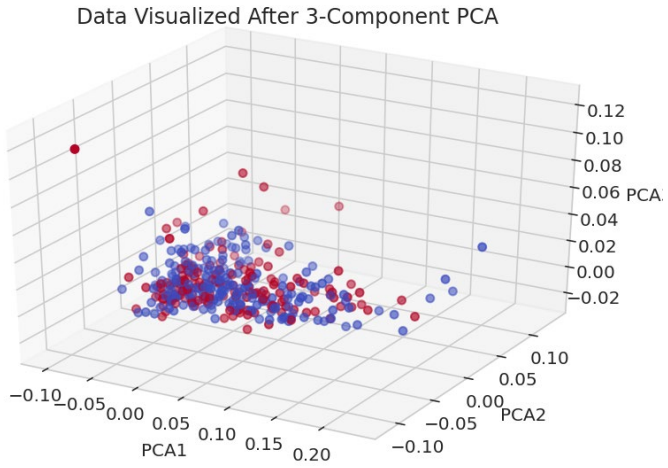

Figure 5: Data Visualized After 3-Component PCA.

ISSN 2616-5791 Vol.4, Issue 3: 25-33, DOI: 10.25236/AJMHS.2023.040305

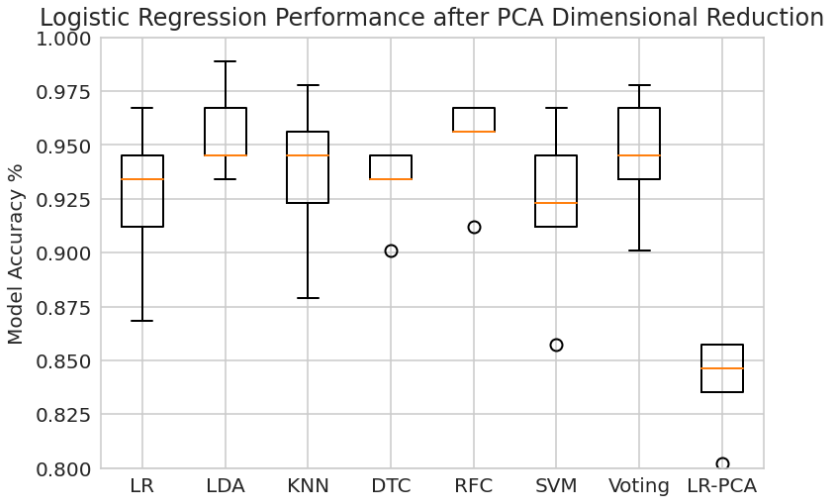

Figure 6: Logistic Regression Performance after PCA Dimensional Reduction.

Figure 6 shows the Logistic Regression Performance after PCA Dimensional Reduction. The overall position of the box line plot of LA-PCA in the figure is much lower than before the dimensionality reduction, and the average accuracy calculated is 0.8395604395604396, which shows that after PCA, the diagnostic LR accuracy is much worse than that of the model without dimensional reduction.

5. Conclusions

In this study, we compiled various data of breast cancer patients, compared the data of malignant tumors with those of benign tumors, created a heat map representing the degree of correlation between the factors using the data of each factor, and used seven machine learning models: Linear Discriminant Analysis, Logistic Regression, K- Nearest Neighbor, Decision Tree Classifier, Random Forest Classifier, Voting Classifier, and Support Vector Machine were used to diagnose breast cancer, and box plots of the accuracy of the corresponding diagnoses were produced for comparison. Subsequently, a dimensionality reduction method was used to reduce 30 features to principal components (PCA) to maximize the explanation of the differences in the data, and the accuracy was calculated and compared again for the reduced models. We conclude that:

1) Among those with breast cancer, the odds of diagnosing a benign tumor is greater than that of a malignant tumor, with a multiplier of about 1.7 for 569 people.

2) Diagnosis was closely related to the mean value of radius, perimeter and concavity, the worst value of radius, perimeter and concavity, and to the mean value of smoothness and symmetry, the standard error of radius and perimeter error, and the worst value of smoothness and symmetry are correlated, and almost independent of the standard error of smoothness, concavity, and symmetry.

3) Radius has a very strong correlation with perimeter, and both of them are independent of smoothness and symmetry.

4) Using these 30 histological tumor characteristics can predict the diagnosis of breast cancer with an accuracy of about 95%.

5) In the accuracy of diagnosing breast cancer, the machine learning algorithms LDA and RFC performed the best and the stability of diagnostic results under large-scale data was also better.

6) After PCA, the average accuracy of LR is worse than that of the model without dimensionality reduction.

References

[1] L. G. Ahmad, A. T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi and A.R. Razavi, "Using three machine learning techniques for predicting breast cancer recurrence," (2013), J Health Med Inform 4: 124. doi: 10.4172/2157-7420.1000124.

ISSN 2616-5791 Vol.4, Issue 3: 25-33, DOI: 10.25236/AJMHS.2023.040305

[2] Uma Ojha and Savita Goel, "A study on prediction of breast cancer recurrence using data mining techniques," 2017 7th Int. Conf. on Cloud Computing, Data Science & Engineering – Confluence, pp 527-530, IEEE, 2017.

[3] Delen D, Walker G, Kadam A (2005) Predicting breast cancer survivability: a comparison of three data mining methods. Artiicial Intelligence in Medicine 34: 113-127.

[4] Mandeep Rana,"Breast cancer diagnosis and recurrence prediction using machine learning techniques", International journal of research in Engineering and Technology, Vol.4, No.4, pp.372-376, April 2015.

[5] A. F. M. Agarap, "On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset," in Proceedings of the 2nd International Conference on Machine Learning and Soft Computing, Phu Quoc Island Viet Nam, Feb. 2018, pp. 5–9. doi: 10. 1145/ 3184066. 3184080.

[6] C. Cortes and V. Vapnik. 1995. Support-vector Networks. Machine Learning 20.3

[7] Medjahed SA, Saadi TA, Benyettou," A. Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules",International Journal of Computer Applications. 2013 Jan 1, vol. 62 (1).

[8] U. K. Kumar, M. B. S. Nikhil, and K. Sumangali, "Prediction of breast cancer using voting classifier technique," in 2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), Chennai, India, Aug. 2017, pp. 108–114. doi: 10.1109/ICSTM.2017.8089135.

[9] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," vol. 23, 2001.

[10] Y. Yasui and X. Wang, Statistical Learning from a Regression Perspective by BERK, R. A., vol. 65, no. 4. 2009.

[11] H. Abbasian, B. Nasersharif, A. Akbari, M. Rahmani and M. S. Moin, "Optimized linear discriminant analysis for extracting robust speech features," 2008 3rd Int.Symp.on Communications, Control and Signal Processing, pp 819-824, IEEE, 2008.

[12] A. F. M. Agarap, "On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset," in Proceedings of the 2nd International Conference on Machine Learning and Soft Computing, Phu Quoc Island Viet Nam, Feb. 2018, pp. 5–9. doi: 10.1145/3184066.3184080.

[13] H. Hasan and N. M. Tahir, "Feature selection of breast cancer based on principal component analysis," in Signal Processing and Its Applications (CSPA), 2010 6th International Colloquium on, 2010, pp. 1-4.