Breast cancer prediction based on multiple machine learning algorithms

Tianyu Sun *

Department of Computer Science and Information Systems, University of Limerick, Limerick, Ireland

* Corresponding author: 23201185@studentmail.ul.ie

Abstract. Breast cancer is one of the common cancers, and its timely and accurate prediction is necessary. In this study, a variety of breast cancer prediction algorithms implemented in Python are studied, compared, and optimized horizontally and vertically. In this study, the problem of breast cancer prediction is deeply studied. By comparing algorithms such as logistic regression, decision tree, random forest, K-Nearest Neighbor (KNN) and support vector Machine (SVM), and introducing two optimization methods, to explore the effective ways to improve the accuracy of breast cancer prediction. It was found that of the five machine learning algorithms provided by the tested sklearn library, logistic regression, random forest and support Vector Machine (SVM) performed well in this breast cancer prediction dataset application. The method of slightly increasing the fitting was calculated and used, the logistic regression statistical fitting method was used and the data was predicted again to obtain better prediction results. Finally, the prediction accuracy of 98.83% was achieved by the optimization method. This provides important guidance for decision makers in the selection of appropriate breast cancer prediction algorithms, which provides stronger support for the early diagnosis and treatment of breast cancer in the future.

Keywords: Breast cancer, Machine learning, Cancer prediction, classification algorithm.

1. Introduction

Breast cancer is the leading cancer in women worldwide. Breast cancer is caused by the abnormal growth of certain cells in the breast. Breast cancer is increasingly becoming the primary cause of death among women globally. Simultaneously, it has been affirmed that ensuring the longevity of patients relies on the confirmed early detection and accurate diagnosis of the disease [1]. Several techniques have been introduced to correctly diagnose breast cancer. Breast screening or mammography is a technique to diagnose breast cancer. It is used to check a woman's condition by X-ray. In general, it is almost impossible to detect breast cancer at the initial stage due to the small size of the cancer cells seen from the outside [2]. Nevertheless, as data accumulates continuously and computing power advances, machine learning algorithms have emerged as a potent tool for conducting predictive analyses in the context of breast cancer [3].

Traditional cancer detection methods are based on the "gold standard" approach and consist of three tests: clinical examination, radiological imaging, and pathological tests [4]. New machine learning techniques and algorithms are designed based on models. The model is designed to predict unseen data and provide good, expected results during its training and testing phases [5]. For example, M Fernandez-Delgado et al. developed 179 classifiers using C and R to test multiple breast cancer datasets and obtained considerable prediction accuracy. Their team's research showed that random forest was the most likely best classifier, achieving 94.1% accuracy. In addition, SVM has also achieved good results [6].

The objective of this study is to examine and compare the performance and applicability of five commonly used machine learning algorithms in the context of prediction problems. These algorithms are logistic regression, decision tree, random forest, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) vector machine. This paper focuses on comprehensive prediction and explores the accuracy of various algorithms for Wisconsin breast cancer data. The goal of this paper is to predict the likelihood of a disease before it occurs [6]. This is to predict the likelihood of cancer-based on the features in the dataset. Through this research, it is hoped that it can help policymakers better

understand the application scope of machine learning algorithms and give more assistance in solving cancer prediction so that they can choose the appropriate algorithm to solve various practical problems more wisely.

The plan of the research presented in this paper is as follows: First, the working principle and mathematical foundation of each machine learning algorithm will be introduced in detail. Then, the datasets and experimental Settings used in this paper are listed. After that, we will show and compare the performance of the algorithms on various prediction problems (using the sklearn library for model building) and make comprehensive predictions. Finally, the research results are summarized, and the advantages and application scenarios of each algorithm are emphasized, as well as the future research directions.

2. Data Source Introduction

2.1. Datasets

The datasets used in this article from the Kaggle website about Wisconsin (diagnosis) of breast cancer data set (https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data). Features are computed from digital images of fine needle aspiration (FNA) of breast masses. They describe the characteristics of the nuclei present in the image [7].

2.2. Methods

2.2.1. Traditional methods

In this paper, the prediction primarily relies on logistic regression, decision trees, random forest, k-nearest neighbor algorithm, and support vector machine (SVM). For each model, the code is written using the sklearn library on the Google Colab platform, citation runs on the same Wisconsin Breast Cancer (diagnosis) dataset, USA, and outputs the predicted confusion matrix along with its accuracy, sensitivity, and specificity. The specific machine learning algorithm is described as follows.

Logistic regression, as a classical statistical learning method, is widely used in binary classification problems. It builds on linear regression and maps the output to a range of (0,1) via a logistic function to accommodate probabilistic problems.

Decision trees, a classification and regression algorithm based on a tree-like structure, make predictions by dividing the input space into different regions.

Random forest, is a method based on ensemble learning, which combines multiple decision trees to improve the accuracy and robustness of the prediction.

The K-Nearest Neighbor (KNN) algorithm, based on example-based learning, classifies the input samples by measuring the distance between the input samples and the known samples. It is a non-parametric method that does not require assumptions about the data distribution and thus performs well when dealing with diverse data.

Support Vector Machine (SVM) is a binary classification model that delineates distinct classes of data by identifying an optimal hyperplane within a high-dimensional space. SVM has strong generalization performance and adaptability to high-dimensional data. (Method)

The performance of these five algorithms on the same US Wisconsin breast cancer dataset and prediction problem will be compared. Their accuracy and robustness will be evaluated and the prediction results will be secondary processed to achieve the purpose of selecting the best of the best. By deeply exploring the strengths and limitations of these machine learning algorithms, it is hoped that data scientists will provide useful information on choosing the right algorithm to meet the needs of different application domains.

2.2.2. Improved model

Since there is still room to improve the results of the previously mentioned models, the paper proposes two new methods that aim to greatly increase the predictive power of the data by slightly increasing the fit.

The explanation of both methods is as follows:

First, predictions are made using the five methods mentioned earlier, and the predictions are combined into a new dataset.

The second thing is to take the mode of the result

In this method, the prediction results of the above five methods are comprehensively predicted, and the final prediction results are given by using the mode, and the relationship between the accuracy and the above algorithms is explored.

Finally, logistic regression on the results

The Results Logistic Regression is the re-prediction of the prediction results of the five algorithms to form the final prediction results. The purpose of this method is to solve the way of using logistic regression, by taking the prediction results of other algorithms as input, the coefficients of five ways as feature values can be obtained, and the final cancer prediction model can be further adjusted and optimized to achieve better performance.

3. Analysis of results

3.1. Result analysis of the traditional model

For the confusion matrix of the five algorithms and their accuracy, sensitivity, and specificity, see Fig. 1, and Fig. 2. In the confusion matrix shown in Figure 1, the Reference Positive entry is the number 0, which means the actual result is malignant, and the Reference Negative entry is the number 1, which means the actual result is benign. For the breast cancer prediction confusion matrix, The highest recognition accuracy should be guaranteed with the lowest False Negative items. It can be seen from the figure that for positive class identification (malignant), logistic regression, random forest, KNN, and SVM have good effects on positive class identification, while the decision tree has a relatively low recognition rate. However, for the negative class recognition (benign), there is little difference between the recognition rates of various algorithms.

For the key False Negative, the number of false predictions in the decision tree method is much higher than that of the other test algorithms. Even if the other speed measurement algorithms have little difference, they still cannot effectively identify the false positive class.

Fig. 2 shows the exact numbers of accuracy, sensitivity, and specificity of each model, and its analysis results are the same as Fig. 1.

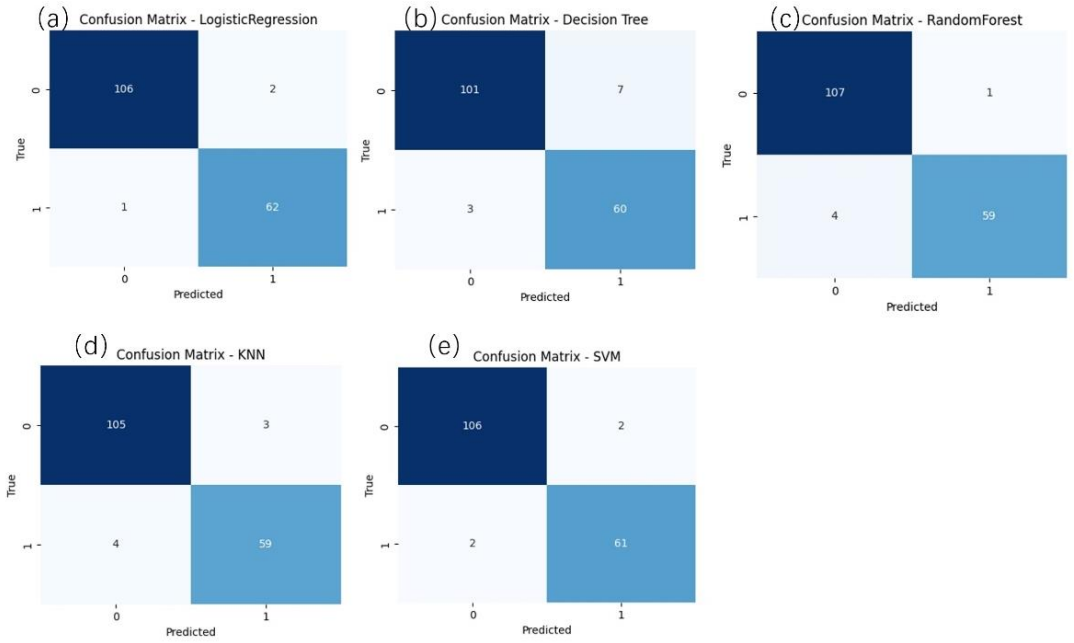

Figure 1. Confusion matrices for logistic regression (a), decision tree (b), random forest(c), KNN(d), and SVM(e)

```
Logistic Regression Metrics:
Confusion Matrix:
[[106   2]
 [  1  62]]
Accuracy: 0.9824561403508771
Sensitivity (True Positive Rate): 0.9841269841269841
Specificity (True Negative Rate): 0.9814814814814815

Decision Tree Metrics:
Confusion Matrix:
[[100   8]
 [  3  60]]
Accuracy: 0.935672514619883
Sensitivity (True Positive Rate): 0.9523809523809523
Specificity (True Negative Rate): 0.9259259259259259

Random Forest Metrics:
Confusion Matrix:
[[107   1]
 [  3  60]]
Accuracy: 0.9766081871345029
Sensitivity (True Positive Rate): 0.9523809523809523
Specificity (True Negative Rate): 0.9907407407407407

SVM Metrics:
Confusion Matrix:
[[106   2]
 [  2  61]]
Accuracy: 0.9766081871345029
Sensitivity (True Positive Rate): 0.9682539682539683
Specificity (True Negative Rate): 0.9814814814814815

KNN Metrics:
Confusion Matrix:
[[105   3]
 [  4  59]]
Accuracy: 0.9590643274853801
Sensitivity (True Positive Rate): 0.9365079365079365
Specificity (True Negative Rate): 0.9722222222222222
```

Figure 2. Accuracy, sensitivity, and specificity of each model

3.2. Results analysis of the improved model

In order to optimize the results in a simpler direction without changing the structure of sklearn's inner prediction model, two re-prediction methods for the prediction results are proposed in the previous part of this paper: taking the mode of the result; the Result Logistic regression.

For the two experimental algorithms, the confusion matrix as well as the accuracy, sensitivity, and specificity are shown in Fig. 3 (final_prediction1: taking modal results, final_prediction2: the Results Logistic Regression). By analyzing Fig.3 and Fig. 4, it can be seen that for positive class recognition (malignant), the accuracy of the two algorithms in identifying the positive class is 0.9682, which is higher than that of the previous test algorithm and at a higher level.

For negative class recognition (benign), there is little gap between the two algorithms, which indicates that there is still room for improvement for negative class recognition.

For the critical False Negative, only one result was successfully identified by taking the mode of the result, while all the positive examples were identified by the Results Logistic Regression, indicating that in this data set, the Results Logistic Regression can successfully meet the criteria mentioned in the beginning of this paper.

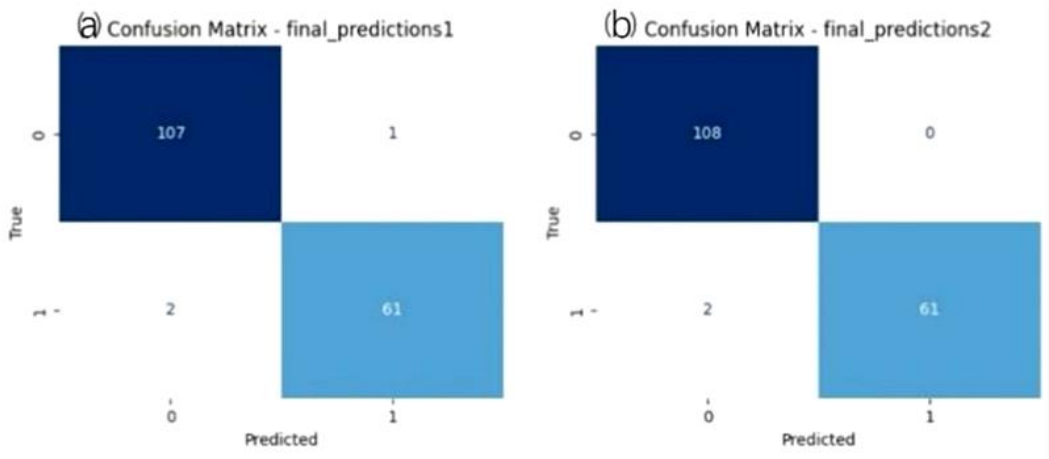

Figure 3. Confusion matrices for the two new algorithms

```
Finall Metrics:
Confusion Matrix:
[[107   1]
 [  2  61]]
Accuracy: 0.9824561403508771
Sensitivity (True Positive Rate): 0.9682539682539683
Specificity (True Negative Rate): 0.9907407407407407


Final Logistic Regression Metrics:
Confusion Matrix:
[[108   0]
 [  2  61]]
Accuracy: 0.9883040935672515
Sensitivity (True Positive Rate): 0.9682539682539683
Specificity (True Negative Rate): 1.0
```

Figure 4. Accuracy, sensitivity, and specificity of the two new algorithms

Table 1. Weight of each sub-algorithm in the Result Logistic Regression algorithm

Model	Logistic regression	Decision tree	Random forest	KNN	SVM
Regression coefficients	2.1616	1.6327	1.1802	1.3179	1.2650
Bias=3.90694					

For outcome logistic regression, the various models are presented in order of importance of the coefficients. Table 1 expresses the weight of the listed algorithms on the Results Logistic Regression in the new algorithm studied in this paper (the larger the number, the greater the weight).

For the two proposed algorithms, the analysis visualization bar graph, where the end Logistic Regression is the Results Logistic Regression, is shown in Figure 5. As can be seen from the figure, the logistic regression model performs well with high accuracy, sensitivity and specificity. It effectively identifies both positive and negative examples. The decision tree model performs well with good accuracy and sensitivity. However, the specificity is slightly lower compared to logistic regression. Random forest models have demonstrated high accuracy, sensitivity, and specificity. It is particularly good at correctly identifying negative examples, and overall optima with decision trees. The SVM model performs strongly overall with high accuracy, sensitivity and specificity. The KNN model performs well with high accuracy and specificity. It is excellent at correctly identifying negative examples, but slightly less sensitive for positive examples.

The Final1 model demonstrated high accuracy, sensitivity, and specificity. It is robust in the recognition of positive examples and high in the accuracy of negative examples, which clearly successfully summarizes and optimizes the five models. The Results Logistic Regression model performed excellently with high accuracy, sensitivity, and specificity. It is robust in the recognition of positive examples while achieving the highest level of accuracy in negative examples. Obviously, for this dataset, among the multiple methods used in this paper, the use of this aggregation method can achieve the highest recognition rate for breast cancer.

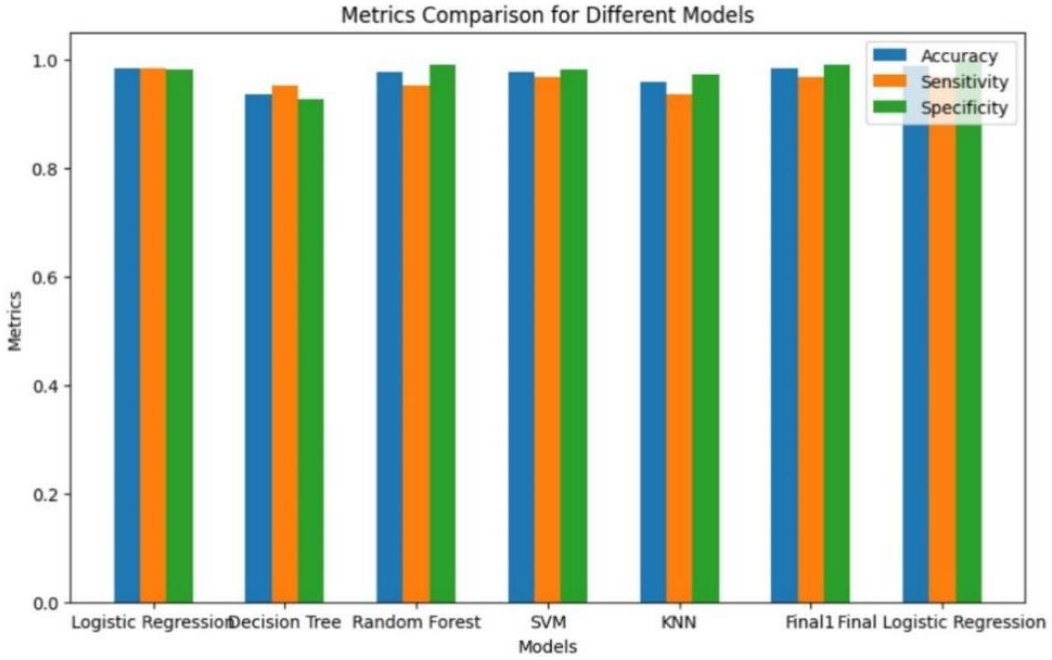

Figure 5. Bar charts of accuracy, sensitivity, and specificity of the enumerated algorithms in this paper

4. Discussion

This study shows that for the five common breast cancer prediction algorithms of logistic regression, decision tree, random forest, KNN and SVM the alignment prediction results are taken mode or the Results Logistic Regression is performed again, and the prediction accuracy of Wisconsin breast cancer dataset is slightly improved. This also proves that processing the result multiple times has a positive effect on improving the prediction accuracy for a fixed dataset.

In conclusion, the Results Logistic Regression algorithm improves the accuracy of the logistic regression algorithm by about 0.6%, reaching 98.83%. At the same time, the accuracy of the negative prediction of this group of data reaches 100%, which indicates that this prediction method can effectively screen out most patients. However, further research is needed to explore the problem of overfitting achieved by this approach.

The accuracy of the algorithm on this dataset is 98.83%. In a subset of papers, Kapil and Rana's improved decision tree technique achieved around 99% accuracy on the WBCD dataset. In the study of Senapati et al, RBFNN-KPSO and RBFNN-extended Kalman filters were used to achieve a classification accuracy of 97.85% and 96.4235%, respectively. In addition, the mathematical model established by Hasan et al through the symbolic regression of polygenic genetic programming performs well, with the highest accuracy of 99.28%. These results demonstrate the potential of different algorithms for breast cancer prediction and remind us that accuracy is affected by many factors, including dataset, feature selection, and algorithm parameters [1].

This puts forward the Results Logistic Regression algorithm, which is a fusion algorithm after all, and cannot help the progress of the machine learning algorithm itself. It aims to provide a more accurate conclusion in predicting breast cancer. However, it is also pointed out that if the basic algorithm of species used in this paper is optimized, and the Results Logistic Regression is used for integration, it will be a very positive help for the accuracy of breast cancer prediction.

5. Conclusion

This study systematically investigated effective ways to improve the accuracy of breast cancer prediction by deeply comparing the performance of five machine learning algorithms, including

logistic regression, decision tree, random forest, k-nearest Neighbor (KNN) and support vector machine (SVM), and introducing two optimization methods. The findings indicate that logistic regression and random forest demonstrate notable performance among conventional machine learning algorithms. Through optimizing the model and employing the results of logistic regression, the prediction accuracy is successfully enhanced to 98.83%. The findings of this study provide more comprehensive guidelines for decision makers to select algorithms for breast cancer prediction, highlighting the superior performance of logistic regression and random forests in prediction. The introduction of optimization methods further improves the accuracy of the algorithm and provides more accurate and reliable support for the early diagnosis and treatment of breast cancer. Future research directions could include more in-depth parameter tuning, optimization of the algorithm structure, and a more detailed analysis of other factors that may affect the prediction results. By continuously improving the comprehensiveness and generalization of the algorithm, we are expected to better understand the complexity of breast cancer and provide more effective decision support for clinical practice. This study not only has important practical application significance in the field of breast cancer but also provides useful experience and enlightenment for the wide application of machine learning in the medical field.

References

[1] El-Baz, A. H. Hybrid intelligent system-based rough set and ensemble classifier for breast cancer diagnosis. Neural Comput & Applic, 2015, 26, 437 – 446.

[2] Islam, M. M., Haque, M.R., Iqbal, H. et al. Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. SN COMPUT. SCI. 2020, 1, 290.

[3] Lu, Y., & Han, J. Cancer classification using gene expression data. Information Systems,2003, 28 (4), 243 - 268.

[4] Vaka, A. R., Soni, B., K., S. R. "Breast cancer detection by leveraging Machine Learning." ICT Express, 2020, 6 (4), 320 - 324.

[5] Bishop, Christopher M. Model-based machine learning. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 2013, 371, 1984.

[6] Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research, 2014, 15 (1), 3133 - 3181.

[7] Cruz, J. A., & Wishart, D. S. Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Informatics, 2016, 2.