EFFECTIVE CLASSIFICATION OF BREAST CANCER USING OUTLIER REMOVAL METHODS AND TRADITIONAL MACHINE LEARNING ALGORITHMS

M.M.Achini Nisansala

Department of ICT, University of Vavuniya

Abstract

Experimental results over the last few years report breast cancer to be the most common type of cancer diagnosed in women's bodies. Though it can arise at any age in a woman's life women over 50 years have a high risk of getting breast cancer. Around 2.3 million new cases are found every year, and among them, around 0.68 million die globally. There are two types of breast tumours: benign and malignant. Diagnosing breast cancer is kind of tough due to the compound nature of the breast cancer cells. However, the treatments for breast cancer are very effective when the disease is diagnosed at an early stage. In this study seven machine learning algorithms were used: Logistic Regression (LR), Linear Discriminant Analysis (LDA), K- Nearest Neighbor (KNN), Gaussian Naive Bayes (GN), Decision Tree Classifier (C4.5), Support Vector Classifier (SVC) and Random Forest (RF) on Wisconsin Breast Cancer Diagnostic Dataset (WBCD) collected from UCI repository for classifying the tumours into benign and malignant. This analysis was carried out in three approaches, without removing the outliers from the dataset, after removing the outliers from the dataset using the interquartile range, and after removing outliers from the dataset using the z-score treatment. Based on the analysis without removing the outliers Logistic Regression (LR) outperformed other classifiers with 95.61% accuracy. After removing the outliers in the interquartile range approach, Gaussian Naive Bayes (GN) achieved the highest accuracy of 97.09%. Z-score outlier treatment reached the highest accuracy of 97.27% among all approaches using the SVC algorithm marking it as the most appropriate method for classifying breast cancer.

Keywords: Breast cancer, outliers, classification, accuracy

[*] *Corresponding Author: nisansala491@gmail.com*

EFFECTIVE CLASSIFICATION OF BREAST CANCER USING OUTLIER REMOVAL METHODS AND TRADITIONAL MACHINE LEARNING ALGORITHMS

M.M.Achini Nisansala
Department of Information and Communication Technology University of Vavuniya
Sri Lanka
nisansala491@gmail.com

INTRODUCTION

The number of patients suffering from cancer diseases has increased rapidly. This makes cancer, the second leading cause of death throughout the world. Among them, the most common type of cancer affecting women is breast cancer (WHO | Breast Cancer, 2021). There are two types of tumours that make suspension of breast cancers: benign tumours and malignant tumours. Benign tumours are considered noncancerous less harmful tumours as they are growing very slowly and do not spread to other parts of the body. However malignant tumours enlarge very fast and they invade and damage healthy tissues and expand throughout the body (Stanford Health Care, 2022). To escalate the survival rate of breast cancer, early accurate detection is the most important thing. In order to detect breast cancer patients have to go through several medical examinations as these tumours are very hard to detect even by specialists in the field. Mammography, biopsy, and ultrasound are some of the common examination types of detecting cancers.

By taking the microscopic image numerical features like area, texture, perimeter, concavity, concave points, and radius of the cells and tissues are calculated. This paper mainly addresses the comparison of the performance and accuracy level between seven machine learning algorithms: Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Gaussian Naive Bayes (GN), Decision Tree (C4.5), Support Vector Classifier (svc) and Random Forest (RF) for specifically determining the benign and malignant tumours. The performance of these seven algorithms on the Wisconsin breast cancer diagnosis dataset (WBCD) is carried out in three procedures:

- Procedure 1 - Taking all the records in the WBCD dataset. As there are no null values available in the dataset dealing with missing data or feature elimination is not required.
- Procedure 2 - Remove the outlier records of the WBCD dataset by taking the interquartile range on the ten most affecting feature sets of the WBCD dataset.
- Procedure 3 - Remove the outlier records of the WBCD dataset by taking the z-score outlier removal from the ten most affecting feature sets of the WBCD dataset.

Outliers are any observations that give some abnormal values that do not fall within the expected distribution of particular data values. On most occasions, there may be some errors that happened while collecting the data while in some occasions these can be due to the varying body structures such as obesity or being skinny. Besides both of these mentioned facts, comparing the accuracy and finding the best approach may help in finding the best method for overall body structures and average-level body structures separately. The accuracy and the performance of these three procedures will be compare and determine the best method to classify benign and malignant tumours.

METHODOLOGY

A. Dataset and attributes

This research paper uses the publicly available dataset Wisconsin Breast Cancer Diagnosis Dataset (WBCD). The dataset consists of 32 attributes including the Id and the diagnosis. Each record consists of two output possibilities: benign or malignant which is included in the diagnosis column. The number of detailed data attributes related to breast cancer is thirty and all these attributes consist of numerical values. These 30 attributes are the average (mean), standard error (SE), and worst, where each attribute consists of texture, radius, area, perimeter, density, smoothness, indentation, concave point, symmetry, and fractal dimensions (Stanford Health Care, 2022). A total of 569 records are available as 357 benign and 212 malignant cases.

B. Experimental environment

All the experiments carried out in this research run on the Jupiter notebook in an anaconda environment. Machine learning models were implemented using the sci-kit-learn (Sklearn) package.

C. Model optimization and training

This research is carried out in three main procedures. The dataset is split into two portions as 80:20 training and testing sets respectively.

- Procedure 1: Without removing any attribute or the record in the dataset all the data available in WBCD dataset are directly used in this approach. The dataset consists of 569 records belonging to 357 benign cases and 212 malignant cases. Scaling of data and hyper-parameter tuning is used for increasing the accuracy of the model.
- Procedure 2: Select the first 10 most influencing fields on the dataset. Figure. 1 gives the most influencing field list. Then remove the outliers in the above 10 fields from the dataset using the interquartile range. After removing the outliers from the dataset number of records available in the dataset was reduced to 518 total records consisting of 354 benign and 164 malignant.
- Procedure 3: Remove the outliers from the 10 most influencing fields. Check for the outliers in the dataset by taking the Z-score method. After removing the outliers from the dataset number of records available in the dataset was reduced to 546 total records consisting of 355 benign and 191 malignant records.

Scaling the data using a min-max scaler is carried out in order to increase algorithm effectiveness and speed up machine learning processing. Hyperparameter tuning is used to optimize the model’s predictive accuracy. Grid search with Cross Validation is used in both approaches for improving the performance of each model. Figure. 2 contains the main steps used in the proposed system of predicting breast cancer.

D. Evaluation of models

These models were evaluated using Accuracy (AC), Precision (PR), Recall (RE), and F1-Score(F1).

RESULTS AND DISCUSSION

After concluding the implementation of machine learning algorithms on the Wisconsin Breast Cancer Diagnostic dataset (WBCD), different performance metrics such as confusion matrix, accuracy, PR, RE, F1-score, and Area Under Curve (AUC) were used for evaluating the performance difference of each of these procedures separately that were followed in this research. Table I summarizes the accuracy obtained by each of the classification algorithms. Based on TABLE I we can find that the best accuracy of 97.82% on the training data set is accomplished by the SVC in procedure 1 and KNN in procedure 2 respectively.

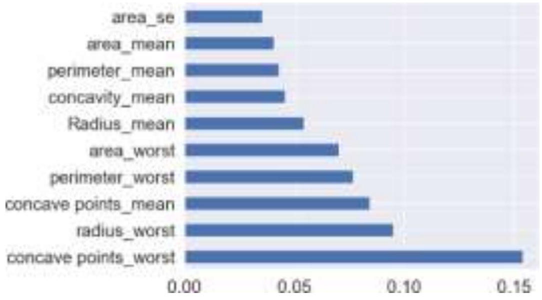

Figure 1: Most influencing fields

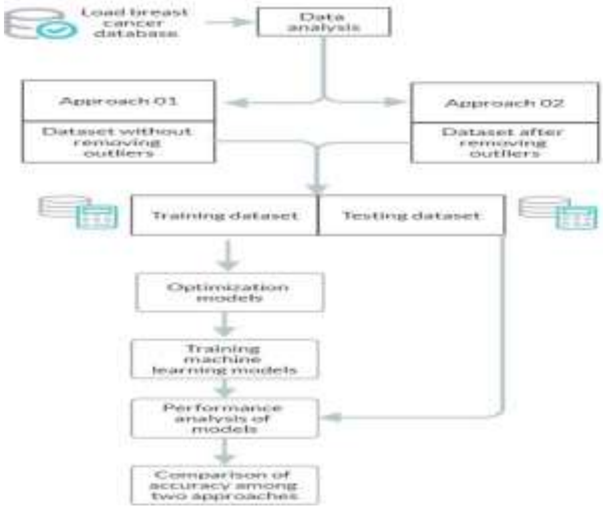
Load breast cancer database
Data analysis
Approach 01
Dataset without removing outliers
Approach 02
Dataset after removing outliers
Training dataset
Testing dataset
Optimization models
Training machine learning models
Performance analysis of models
Comparison of accuracy among two approaches

Figure 2: Methodology

CONCLUSIONS/RECOMMENDATIONS

The application of machine learning algorithms on medical datasets and finding the best methods of diagnosing different types of diseases, predicting the severity range, predicting mere reasons for the diseases, and predicting the condition of the disease are very significant works. They can help in increasing the lifetime of a patient and early detection of the disease can help in gaining the proper treatments on time before making it into the critical stage. Proper diagnosis of disease and the severity of the disease is very complicated even for experienced professionals.

In this work, we have used the Wisconsin breast cancer diagnostic dataset (WBCD) and applied 7 machine learning algorithms in three different procedures, and compared and evaluated different results obtained based on confusion matrix, accuracy, precision, recall, and f1-score. After comparing these results in three approaches, we found that the SVC gives the best accuracy on the testing data only after removing the outliers using the z-score treatment mechanism.

In future work, we can apply more machine learning algorithms in more than one dataset related to one disease and can compare and increase the performance of the models.

Table 1.Accuracy of datasets on scaled data

Approach	Algorithm	Training accuracy	Testing accuracy
Procedure 1	LR	96.94%	95.61%
	LDA	95.61%	87.88%
	KNN	97.15%	93.86%
	GN	94.29%	93.86%
	C4.5	91.43%	92.12%
	SVC	97.82%	94.77%
	RF	95.16%	93.03%
Procedure 2	LR	96.13%	95.09%
	LDA	95.15%	92.27%
	KNN	97.82%	94.27%
	GN	92.48%	97.09%
	C4.5	92.26%	87.45%
	SVC	97.33%	97.00%
	RF	96.36%	94.18%
Procedure 3	LR	95.86 %	93.64%
	LDA	94.71%	93.36%
	KNN	97.01%	95.45%
	GN	93.12%	96.36%
	C4.5	92.66%	91.82%
	SVC	97.68%	97.27%
	RF	95.41%	95.45%

REFERENCES

Journal Articles

Aamir, S., Rahim, A., Aamir, Z., Abbasi, S. F., Khan, M. S., Alhaisoni, M., Khan, M. A., Khan, K., & Ahmad, J. (2022). Predicting breast cancer leveraging supervised machine learning techniques. Computational and Mathematical Methods in Medicine, 2022, 1–13. https://doi.org/10.1155/2022/5869529

Ming, C., Viassolo, V., Probst-Hensch, N., Chappuis, P. O., Dinov, I. D., & Katapodi, M. C. (2019). Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models. Breast Cancer Research, 21(1). https://doi.org/10.1186/s13058-019-1158-4

Nasien, D., Enjeslina, V., Adiya, M. H., & Baharum, Z. (2022). Breast Cancer Prediction Using Artificial Neural Networks Back Propagation Method. Journal of Physics, 2319(1), 012025. https://doi.org/10.1088/1742-6596/2319/1/012025

Yazdani, A., Kazemi-Arpanahi, H., Ghalibaf, M. B., & Orooji, A. (2022). Performance evaluation of machine learning for breast cancer diagnosis: A case study. Informatics in Medicine Unlocked, 31, 101009. https://doi.org/10.1016/j.imu.2022.101009

Conferences and proceedings

Ara, S., Das, A., & Dey, A. K. (2021). Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms. 2021 International Conference on Artificial Intelligence (ICAI). https://doi.org/10.1109/icai52203.2021.9445249

Khourdifi, Y., & Bahaj, M. (2018). Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification. 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS). https://doi.org/10.1109/icecocs.2018.8610632

Cancer cells in lymph nodes crib an ‘all access’ pass to metastasize, Stanford researchers find. (2022, August 16). News Center. https://med.stanford.edu/news/all-news/2022/08/cancer-cells-lymph-nodes.html

Naji, M. A., Filali, S. E., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021). Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. Procedia Computer Science, 191, 487–492. https://doi.org/10.1016/j.procs.2021.07.062

Naji, M. A., Filali, S. E., Bouhlal, M., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021). Breast Cancer Prediction and Diagnosis through a New Approach based on Majority Voting Ensemble Classifier. Procedia Computer Science, 191, 481–486. https://doi.org/10.1016/j.procs.2021.07.061

Rabiei, R. (2022). Prediction of Breast Cancer using Machine Learning Approaches. Journal of Biomedical Physics & Engineering, 12(3). https://doi.org/10.31661/jbpe.v0i0.2109-1403

Saleh, H., Ghany, S. F. A., Alyami, H., & Alosaimi, W. (2022). Predicting breast cancer based on optimized deep learning approach. Computational Intelligence and Neuroscience, 2022, 1–11. https://doi.org/10.1155/2022/1820777

Saranya, S., & Sasikala, S. (2020). Diagnosis using data mining algorithms for malignant breast cancer cell detection. 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). https://doi.org/10.1109/iceca49313.2020.9297481

Soria, D., Garibaldi, J. M., Biganzoli, E., & Ellis, I. O. (2008). A Comparison of Three Different Methods for Classification of Breast Cancer Data. 2008 Seventh International Conference on Machine Learning and Applications. https://doi.org/10.1109/icmla.2008.97

Web sources

Stanford Health Care. (2022). (Stanford Medicine) Retrieved from https://stanfordhealthcare.org/medical-conditions/cancer/cancer.html

WHO | Breast Cancer. (2021, March 26). (WHO) Retrieved from https://www.who.int/news-room/fact-sheets/detail/breast-cancer

Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (n.d.). breast-cancer-wisconsis-data. Retrieved from UCI.