17th Iranian Int'l Industrial Engineering Conference
هفدهمین کنفرانس بین المللی مهندسی صنایع
3 & 4 February 2021 / Ferdowsi University of Mashhad
۱۵ و ۱۶ بهمن‌ماه ۱۳۹۹ / دانشگاه فردوسی مشهد

A comparison between different classification algorithms for predicting metastasis in breast cancer patients

Payam Mahmoudi[1], Arman Behnam[2], Toktam Khatibi[3]
[1]Industirial Engineering Department, Iran University of Science and Technology; payam_mahmoudi97@ind.iust.ac.ir
[2]Industrial Engineering Department, Iran University of science and technology; arman_behnam@ind.iust.ac.ir
[3] Assistant Professor, Industrial & Systems Engineering Department, Tarbiat Modares University, Tehran, Iran; toktam.khatibi@modares.ac.ir
* Corresponding author: Toktam Khatibi

ABSTRACT
Breast cancer is one of the most common cancers among women around the world. According to World Health Organization (WHO), breast cancer is second reason for cancer mortality. Approximately 30%-40% patients suffering from breast cancer will experience recurrence and 10%-15% of them were reported to die of cancer metastasis. Early diagnosis or prediction of metastasis will reduce mortality rate and treatment cost. In this study we have used a data set containing 555 record of patients with breast cancer (83 have experienced metastasis) and 8 features. Several machine Learning algorithms including Random Forest (RF), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Multi-Layer Perceptron (MLP) were used to predict metastasis. Total accuracy and area under curve (AUC) extracted out of Receiver operating characteristic values were used to evaluate models. The results show that Multi-Layer Perceptron Outperform other methods to predict the metastasis.
Keywords: Breast Cancer, Metastasis, Prediction, Machine Learning, Deep Learning

1. Introduction

Breast cancer is one of the most common cancers among women around the word. According to World Health Organization (WHO) breast cancer is second reason for cancer mortality [1]. Approximately 30%-40% patients suffering from breast cancer will experience recurrence [2] and 10%-15% of them were reported to die of cancer metastasis [3]. Early detection of breast cancer and its recurrence can really help to prevent death, reduce mortality rate and treatment cost. With early diagnosis of breast cancer, 97% of women survive for 5 years or more [4].

Machine learning (ML) techniques can help us to improve diagnostic capability. With ML diagnostic errors can be reduced which helps to analyze a greater number of breast cancer patients faster. ML consists of various methods. Each method has its own purposes, advantages and disadvantages. There are two kinds of ML algorithms: 1) Supervised Learning 2) Unsupervised Learning. In Supervised learning, the machine will be trained using labeled data and it is used for predicting the new datasets output. Classification and regression are two kinds of supervised algorithms. Unsupervised learning will help to find all kind of unknown patterns in data and describe it. Clustering and association rules are two types of unsupervised learning.

By using classification algorithms, a model from a set of training data whose target class labels are known can be constructed. In recent years many studies have been done regarding to application of machine learning and deep learning algorithms in early diagnosis or prediction of breast cancer. Israni has used PCA as a dimension reduction tool to augment accuracy of diagnosis. SVM, Random Forest, KNN, Decision Tree, NB and ANN were used for creating different models. In conclusion SVM was the superior algorithm [5]. Adel et al have used elastography images for predicting breast cancer. In the preprocess part they have used PCA and finally a SVM model was built which it has sufficient accuracy [6]. Asri have compared SVM, C4.5, NB and KNN algorithms to predict breast cancer [7]. Bataineh has compared machine learning and deep learning algorithms for diagnosis whether the tumor is benign or malignant. Multi-layer perceptron, KNN and regression were used [8].

There are also some studies about predicting or diagnosis of metastasis in breast cancer. In [9] they wanted to predict metastasis in breast cancer. Expectation Maximization were used for imputing missing values and C4.5, SVM and Back Propagation Neural Network algorithms were evaluated.

BPNN was the best model for predicting metastasis in breast cancer. The goal of Zemmour et al was to predict early breast cancer metastasis from DNA microarray data using Elastic Net, LASSO and CosxBoost algorithms [10]. Chen et al predicted chemo-brain in breast cancer survivors using multiple MRI features. They have compared 9 classification machine learning models. LR was the best model for predicting chemo-brain [11]. Tapak et al wanted to predict metastasis and death in breast cancer patients. NB, Random Forest, AdaBoost, SVM, LR and LDA were used to predict [12]. Tseng et al have used Random Forest, SVM, LR and Bayesian Network to predict metastasis in breast cancer patients [13]. Fan et al have used Deep Convolutional Neural Network (CNN) with help of image processing. They have used AUC for evaluating the model [14]. Zhou et al have used 2 independent datasets from 2 different hospitals. CNN algorithm was built to predict lymph node metastasis from primary breast cancer [15]. Sun et al have used ultrasound images to build three CNN algorithms and three radiomics models using Random Forest for predicting axillary lymph node metastasis of breast cancer [16].

In this study, we aim on predicting metastasis with ML-based algorithms to prevail discussed methods and increase accuracy due to sensitivity of the topic with considering people lives. We will use RF, LDA, QDA, SVM, KNN and MLP and compare their results to find the best algorithm describing our data. We implemented substantial tools used in the classification tasks and appropriate metrics such as accuracy and AUC to analyze encounters obstacles in fields of explaining results more meaningfully. Findings of this study can help decision makers to find the positive metastasis cases using our predictions.

2. Materials and methods

2.1. Dataset

We used a dataset originates from a study that was conducted in 2014 in Tehran. Information of patients who suffered from breast cancer and registered with the Comprehensive Cancer Control Center associated with ShahidBeheshti University of Medical Science from 1998 to 2013 is exploited. The dataset contains 555 breast cancer patients. The output of our model is whether a breast cancer patient have suffered from metastasis or haven't (1 or 0). The distribution of patients is shown in figure 1 based on their output value. All patients diagnosed pathologically and patients with unknown pathology were excluded from analysis. We selected 8 risk factors to predict metastasis in patients and compare different classification algorithms. Table 1 shows 8 selected risk factors known as our problem variables.

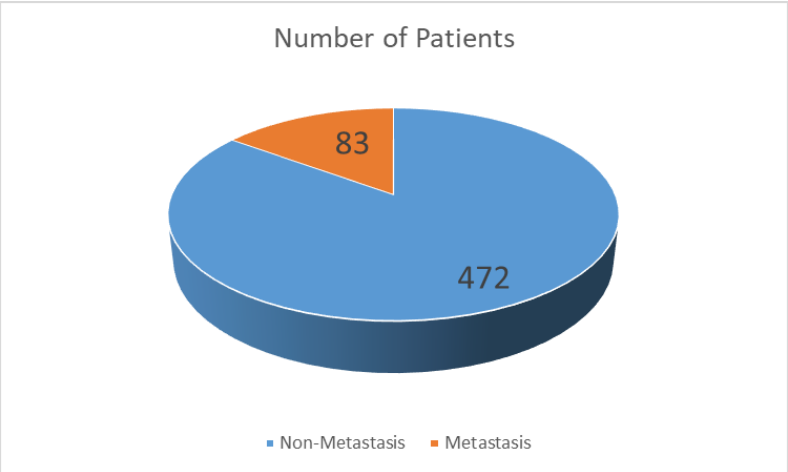

Figure 1. Distribution of 555 patients based on their metastasis situation

Table 1. Risk factors

Factor	Details
Age	Between 24 - 89
Grade	Well, modern and poor
Stage	Stage 1 – Stage 2 – Stage 3 – Stage 4
Estrogen Receptor	Negative / Positive
Progesterone Receptor	Negative / Positive
Human Epidermal Growth Factor Receptor 2	Negative / Positive
Pathological Type	Ductal/lobular carcinoma in situ - Invasive lobular carcinoma – Invasive ductal surgery
Surgical Approach	Modified Radical Mastectomy – Breast-Conserving surgery

2.2.Methods

In this study Random Forest, LDA, QDA, SVM, KNN and MLP algorithms have been implemented to predict metastasis. All of these methods are implemented using Stratified K-fold cross validation for optimizing our answers.

2.2.1. Random Forest (RF)

Random Forest is a supervised ensemble algorithm which is mainly used for classification. This algorithm creates several decision trees on data samples and gets the result of the prediction from each of them. At the end, it selects the best solution by means of voting. This algorithm is better that single decision tree because it uses the average of the results and reduces the over-fitting. RF can find the most important predictors using mean decrease Gini and mean decrease accuracy [22]. Also for unbalanced datasets, it can balance the error and if a large part of the features is lost, it can maintain accuracy.

2.2.2. Linear Discriminant Analysis (LDA)

LDA is a dimensionality reduction technique and it's mostly used for the supervised classification problems, especially for projecting the features in higher dimension space into lower dimension space. LDA creates a linear combination of predictors and classifies the outcome [23]. LDA assumes that each observation is drawn from multivariate Gaussian distribution and predictor variables have common variance.

2.2.3. Quadratic Discriminant Analysis (QDA)

QDA algorithm works similar to LDA. The difference is that QDA assumes that predictor values have different variance.

2.2.4. Support Vector Machine (SVM)

SVM is a powerful kernel-based ML technique for supervised data classification. The basic idea is to create a hyperplane as the decision surface for classification [18]. It can be used for classification and regression approaches. SVM is excellent for non-linear problems [19]. SVM perform its tasks by maximizing the margin separating 2 classes while minimizing the classification errors [20]. SVM can work well on smaller and cleaner datasets and it is less effective on classifying datasets with overlapping

classes.

2.2.5. K-Nearest Neighbors (KNN)

KNN is a supervised machine learning algorithm which is a lazy learner and classifies the dataset based on their similarity. each record is classified by a majority of its neighbors. The neighbors are selected from a training set of records which their classes are known [21]. KNN is easy to implement and doesn't work well with large dataset. It is also sensitive to any noise, missing values and outliers in the dataset.

2.2.6. Multi-Layer Perceptron (MLP)

Multilayer perceptron is defined as a biologically inspired feed-forward network which consists of multiple layers, each containing multiple artificial neuron units and can be used for classification and regression tasks in a supervised learning approach [24].

2.2.7. Stratified K-fold cross validation

Distribution-balanced stratified cross-validation (DBSCV) improves the estimation quality by providing balanced intraclass distributions when partitioning a data set into multiple folds. DBSCV performs better (has smaller biases) than the regular stratified cross validation in most cases, especially when the number of folds is small [25].

3. Results

The patients with breast cancer at diagnosis aged 52.58 year in average with a minimum and maximum of 24 and 89 years respectively. 53% of patients were presented with grade of modern and 42% were at stage 2 and 71%, 68%, 76% had ER+, PR+ and HER2-. 90% of patients were diagnosed with pathological type of invasive ductal carcinoma and 65% received breast-conserving surgery. The correlation between variables are shown in Figure 2.

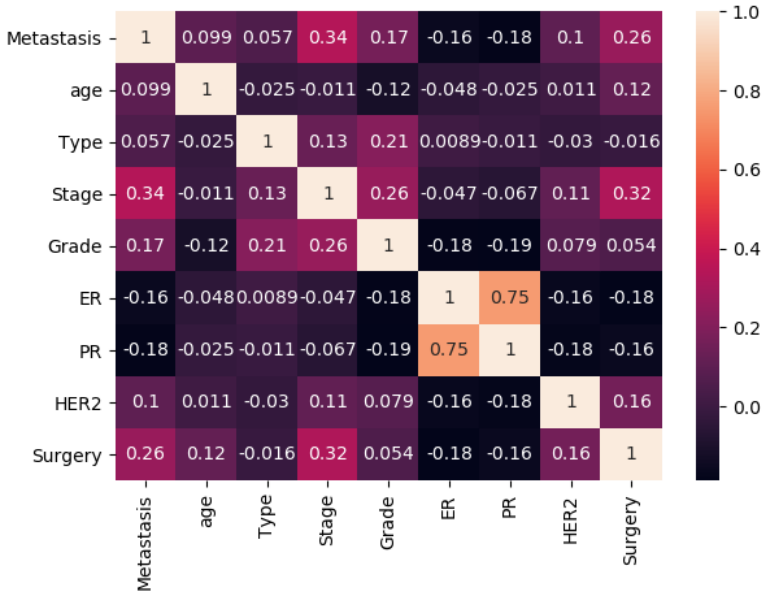

Figure 2. Correlation between variables

The correlation between ER and PR is 0.75 so we can assume that ER and PR are showing the same thing and one of them is sufficient for modeling. We have chosen PR because it has more effect on output. Stage of the tumor has the most effect on metastasis.

We divided the data into two sets of training and testing set. 20% for train and 80% for test. Then Stratified K-fold cross validation is used to optimize our results and resolving the overfitting problem

just in case. Due to our experiment, K has been chosen to be 6 (due to best description of data). The results of applying different classification algorithms are shown in table 2.

Table 2. Algorithms and their performance

Algorithm	Accuracy	AUC
Random Forest	85%	72%
Linear Discriminant Analysis	86%	83%
Quadratic Discriminant Analysis	83%	79%
Support Vector Machine	86%	73%
K-Nearest Neighbors	86%	62%
Multi-Layer Perceptron	91%	76%

According to small amount of data we used in the article, LDA method has the highest AUC value since in such cases LDA has the best variability description due to variables number and rows. For larger datasets, MLP will always have higher AUC value than other ML algorithms because of its neural based nature which is more suitable for more complex cases. Multi-layer perceptron with 84% has the highest accuracy for predicting metastasis which shows how powerful ANNs are in predicting accurate in comparison with other ML methods. Receiver operating characteristic curve and extracted AUC for each fold of dataset for different algorithms are shown in figure 3.

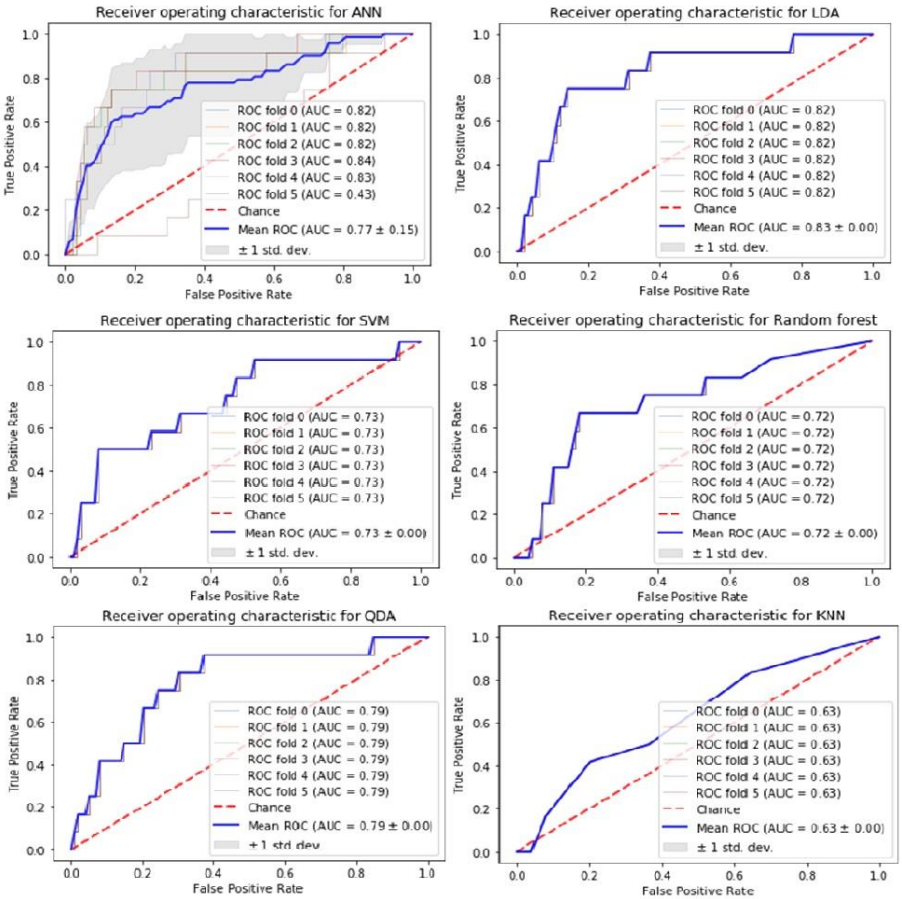

Figure 3. ROC for different algorithms

4. Discussion

In this paper we have studied relevant articles and tried to complete their works with using more different classification algorithms including MLP to have better prediction about metastasis in breast cancer patients.

For future research, more variables should to be considered and many other ML methods may have higher accuracy and better output in forecast procedures. In addition, with a larger dataset, predictions will do better on the dataset. All health decision-making body organizations can rely on our analysis to plan for healthcare programs and also take advantage of this situation to improve conditions.

5. Conclusion

The purpose of this study was to evaluate the performance of six classification algorithms in predicting metastasis occurrence in patients with breast cancer. Our results showed that MLP has the best performance in predicting metastasis with respect to accuracy and AUC. The next best algorithm is LDA. Further investigation is needed using large data sets to recommend a stronger tool for predicting metastasis occurrence in breast cancer patients.

Acknowledgments

There is no financial support for this study. Python software has been used for implementation. Thanks to Leili Tapak from department of biostatistics, school of public health, Hamadan university of medical sciences, Hamadan, Iran who helped us with the preparation of the data.

References

[1] World Health Organization. Breast cancer: breast cancer and early diagnosis. Available from: URL: http://wwwwhoint/cancer/prevention/diagnosis-screening/breast-cancer/en/. 2018.

[2] S.E.Moody, D.Perez, T.C.Pan, C.J.Sarkisian, C.P.Portocarrero, C.J.Sterner, K.L.Notorfrancesco, R.D.Cardiff, L.A.Chodosh, The transcriptional repressor Snail promotes mammary tumor recurrence, Cancer Cell. 8 (2005) 197–209. doi:10.1016/j.ccr.2005.07.009.

[3] L.Cheng, M.D.Swartz, H.Zhao, a. S.Kapadia, D.Lai, P.J.Rowan, T.a.Buchholz, S.H.Giordano, Hazard of Recurrence among Women after Primary Breast Cancer Treatment--A 10-Year Follow-up Using Data from SEERMedicare, Cancer Epidemiol. Biomarkers Prev.21(2012)800–809. doi:10.1158/1055-9965.EPI-11-1089.

[4] Delen, D., et al. (2005). "Predicting breast cancer survivability: a comparison of three data mining methods." Artif Intell Med 34(2): 113-127.

[5] Priyanka Israni (2019). "Breast Cancer Diagnosis (BCD) Model Using Machine Learning." International Journal of Innovative Technology and Exploring Engineering 8(10): 4456-4463.

[6] Adel, M., et al. (2019). Breast Cancer Diagnosis Using Image Processing and Machine Learning for Elastography Images. 2019 8th International Conference on Modern Circuits and Systems Technologies (MOCAST). Thessaloniki, Greece, Greece, IEEE.

[7] Asri, H., et al. (2016). "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis." Procedia Computer Science 83: 1064-1069.

[8] Bataineh, A. A. (2019). "A Comparative Analysis of Nonlinear Machine Learning Algorithms for Breast Cancer Detection." International Journal of Machine Learning and Computing 9(3): 248-254

[9] Lg, A. and E. At (2013). "Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence." Journal of Health & Medical Informatics 04(02).

[10] Zemmour, C., et al. (2015). "Prediction of Early Breast Cancer Metastasis from DNA Microarray Data Using High-Dimensional Cox Regression Models." Cancer Informatics 14s2.

[11] Chen, V. C., et al. (2019). "Predicting chemo-brain in breast cancer survivors using multiple MRI features and machine-learning." Magn Reson Med 81(5): 3304-3313.

[12] Tapak, L., et al. (2019). "Prediction of survival and metastasis in breast cancer patients using machine learning classifiers." Clinical Epidemiology and Global Health 7(3): 293-299.

[13] Tseng, Y. J., et al. (2019). "Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies." Int J Med Inform 128: 79-86.

[14] Fan, K., et al. (2019). Deep Learning for Detecting Breast Cancer Metastases on WSI. Innovation in Medicine and Healthcare Systems, and Multimedia: 137-145.

[15] Zhou, L.-Q., et al. (2020). "Lymph Node Metastasis Prediction from Primary Breast Cancer US Images Using Deep Learning." Radiology 294(1): 19-28.

[16] Sun, Q., et al. (2020). "Deep Learning vs. Radiomics for Predicting Axillary Lymph Node Metastasis of Breast Cancer Using Ultrasound Images: Don't Forget the Peritumoral Region." Frontiers in Oncology 10.

[17] Yang, X., et al. (2020). "Deep Learning Signature Based on Staging CT for Preoperative Prediction of Sentinel Lymph Node Metastasis in Breast Cancer." Academic Radiology 27(9): 1226-1233.

[18] E. Byvatov, et al., Comparison of support vector machine and artificial neural network systems for drug/nondrug classification, J. Chem. Inform. Comput. Sci.43 (6) (2003) 1882–1889.

[19] Yadav, R., et al. (2013). "Chemotherapy Prediction of Cancer Patient by using Data Mining Techniques." International Journal of Computer Applications 76(10): 28-31.

[20] Joachims T (1998) Making large-scale support vector machine learning practical. Advances in Kernel Methods: Support Vector Learning. MIT Press, Cambridge, MA, 169-184.

[21] Kumar, G. R., et al. (2013). "An Efficient Prediction of Breast Cancer Data using Data Mining Techniques." International Journal of Innovations in Engineering and Technology (IJIET) 2(4):139-144

[22] Breiman L. Random forests. Machine learning. 2001;45(1):5-32.

[23] Izenman AJ. Linear discriminant analysis. Modern multivariate statistical techniques: Springer;2013. p. 237-80.

[24] Yan, H., et al. (2006). "A multilayer perceptron-based medical decision support system for heart disease diagnosis." Expert Systems with Applications 30(2): 272-281.

[25] Zeng, X. and T. R. Martinez (2000). "Distribution-balanced stratified cross-validation for accuracy estimation." Journal of Experimental & Theoretical Artificial Intelligence 12(1): 1-12.