Artificial & Computational Intelligence: Published Online (December 2021)
https://acors.org/ijacoi/VOL2_ISSUE4_01.pdf

Artificial & Computational Intelligence

Breast Cancer Prediction using Machine Learning

Shubhangi P, Tanya S, Vandana N
Department of Electronics and Communication Engineering
shubhangi045btece18@igdtuw.ac.in, tanya068btece18@igdtuw.ac.in, vandana7379@gmail.com
Indira Gandhi Delhi Technical University for Women, Delhi, India

Abstract

Breast cancer, a highly common type of cancer, has become a significant public health concern today. According to world statistics, it accounts for most emerging cancer cases and cancer-related deaths today. Importance of diagnosis in people's lives consequently calls for the most effective and accurate cancer predictions, for updating patient treatment aspects and survival criteria.Machine Learning algorithms are widely utilizedin intelligent human services frameworks, particularly for breast cancer diagnosis [4].It has the potential to be extremely useful in the early identification and prediction of breast cancer. It has become a research hotspot and has proven to be a powerful technology. We used machine learning algorithms in this project to predict and diagnose breast cancer by determining the most effective classifier in terms of confusion matrix, accuracy, and precision. The Random Forest classifier outperformed all other classifiers and achieved the greatest accuracy (96.5%). The entire project is carried out in the Anaconda environment and is based on the Python programming language, NumPy, Pandas, Matplotlib, Seaborn, and Scikitlearn libraries.

Keywords: Breast cancer, prediction analysis, machine learning classifiers, confusion matrix, accuracy

1. Introduction

Breast cancer is one of the most lethal and heterogeneous disease in this present era that causes the death of enormous number of women all over the world [3]. It is the second largest disease that is responsible of women death [10]. Timely clinical treatment of patients can significantly improve prognosis and chances of survival, making early diagnosis of breast cancer very important. More accurate classification of benign tumors can help patients in avoiding unnecessary treatments. Therefore, the correct diagnosis of breast cancer and the division of patients into benign or malignant categories has been the subject of many studies. Machine learning is widely recognized as the best method for classifying and predictive modeling of breast cancer patterns because of its special advantage of identifying key features from complex breast cancer datasets.

Data mining algorithms used in the healthcare industry play a key role in disease prediction, diagnosis, drug cost reduction, and real-time life-saving decision-making for high performance.The most common goals of data mining modeling are classification and forecasting[7]. It uses multiple algorithms to predict breast cancer. This project primarily compares the performance of three classifiers:[1] Logistic Regression, Decision Tree and Random Forest. The research community believes that these are some of the most effective data mining algorithms, leaving a significant mark in today's world. Our goal is to use machine learning algorithms to predict and diagnose breast cancers and find the most effective classifier [2] based on the performance of each classifier in terms of confusion-matrix, accuracy, and precision.

The dataset used in this story is publicly available and was created by Dr. William H. Wolberg, physician at the University of Wisconsin Hospital at Madison, Wisconsin, USA.The Wisconsin Breast Cancer dataset is obtained from a prominent machine learning database named UCI machine learning database [6].To create the dataset Dr. Wolberg used fluid samples, taken from patients with solid breast masses and an easy-to-use graphical computer program called Xcyt, which is capable of perform the analysis of cytological features based on a digital scan (Building a Simple Machine Learning Model on Breast Cancer Data, Vishabh Goel, 2018).

2. Methodology

2.1 *Importing the libraries*

Libraries NumPy, pandas, seaborn, matplotlib, and sci-kit learn are imported.

2.2 *Loading the data*

- The Wisconsin dataset is loaded using 'read_csv' method from the pandas library.
- First seven rows are inspected.

2.3 *Exploring the data*

- We get the dimension of the dataset, i.e., the number of rows and columns by the 'shape' method.
- We also get a count of the empty values present in the dataset.

2.4 *Cleaning the data*

- From this step onwards, we start to 'pre-process' the data.Data pre-processing utilizesstrategies like theevacuation of loudinformation, expulsion of missing information, filling default esteems if material and so forth are utilized during revision of information before bolstering it to the calculation [2].
- We determine the rows and columns that contain missing values and drop them since they add no value.
- We get a count of the M's (Malignant) and B's (Benign) cells mentioned in the 'diagnosis' column. Further, we visualize the same using the 'countplot' method from the seaborn library.
- The categorical values (M's and B's) are encoded into numeric data. These numeric values later serve as classifier input. 'M' and 'B' are changed to 1 and 0 respectively.
- To compare all 569 rows within the dataset using column index 1 to 4, we create pairplots using the method 'pairplot' method from the seaborn library.
- The cleaner dataset is re-inspected.
- We get the correlation between the columns and then visualize it using 'corr' and 'heatmap' methods for further exploration.

2.5 *Splitting the dataset into independent(X) and dependent(Y) datasets*

- X has the features that can help us detect cancerous cells.
- Y is for diagnosis, i.e., it tells us whether they are cancerous cells or not.

2.6 *Creating the training and testing dataset*

- Both X and Y are split into 75% training data and 25% testing data using 'train_test_split' method from the Sci-kit library.
- X_train is the 75% which correlates with Y_train. Similarly, X_test is the 25% and Y_test correlates with that.
- We scale or standardize the dataset to bring the values within a certain range.Since most of the machine learning algorithms use Euclidian distance between two data points in their computations [8],we bring all features to the same level of magnitudesto keep them from functioning abnormally which usually happens when the individual features don't resemble the standard normally distributed data.

$$(Value - Mean) / Standard\ deviation$$

2.7 *Creating a function to hold the three different classifier model*

Logistic Regression, Decision Tree and Random Forest classifier – these are the models that will detectthe presence of cancer cells in a patient. Through this function we will also print the accuracy of each model using the training data.

2.8 *Creating the model*

Classification is one of the most important and essential tasks in machine learning and data mining. About a lot of research has been conducted to apply data mining and machine learning on different medical datasets to classify Breast Cancer. Many of them show good classification accuracy [5].

This model will consist of all the models. Further, it will produce the accuracy scores on the training data for each model to classify if the patient has cancer or not.

2.9 *Testing the model*

- We obtain the confusion matrix and the accuracy of each model respectivelyusing the test data.
- The confusion matrix shows us how many patients each model misdiagnosed(number of cancer patients that were misdiagnosed as not having cancer i.e., false negative, and the number of safe patients that were misdiagnosed with having cancer i.e., false positive) and the number of correct diagnoses represented by the true positives and true negatives.
    1. False Positive (FP) = A test result that erroneously indicates the existence of a particular condition or property.
    2. True Positive (TP) = A test result that correctly identifies the presence of a particular value, condition, or property. Sensitivity refers to the measure of the proportion of these true positive values. It is also known as'true positive rate' or 'probability of detection' in certain fields.
    3. True Negative (TN) = A test result that correctly identifies the negative values. Specificity measures the proportion of these true negative values.It is also known as 'true negative rate'.
    4. False Negative (FN) = A test result that shows that a condition is not met when it is met.

		Prediction	
		0	1
Actual	0	TN	FP
	1	FN	TP

Figure 1 shows Confusion Matrix table

2.10 *Printing the prediction of the most accurate classifier and comparing with actual dataset.*

We observe that most of the values in the actual dataset match with the corresponding value in the predicted dataset.

```
[1 0 0 0 0 0 0 0 0 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 0 0 1 0 0 1 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 0 0 1 0 1 1 1 1 0 0 1 0 0
 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 1 0 1 1 0
 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 1]

[1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 1 1 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 1 0 1 0 0 1 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 0 0 1 0 1 1 1 1 0 0 1 0 1
 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 1 0 1 1 0
 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 1]
```

Figure 2 shows thePredicted dataset values Versus Original dataset values

3. Results and Discussion

We observe that the Random Forest classifier outperformed all other classifiers considered in the project with an accuracy of 96.5%. So, I would advise using this model for detecting cancer cells in patients.The model misdiagnosed a few patients as having cancer when they didn't and vice-versa. Models like these are worked upon and improved till they achieve an accuracy of almost 100%, as these models deal with human lives, and even a minute compromise can eventually lead to heavy losses.Machine learning algorithms can be used for medical oriented research as it advances the system, reduces human errors, and lowers manual mistakes [9].

```
Model  0
[[86  4]
 [ 3 50]]
Testing accuracy =  0.951048951048951

Model  1
[[83  7]
 [ 2 51]]
Testing accuracy =  0.9370629370629371

Model  2
[[87  3]
 [ 2 51]]
Testing accuracy =  0.965034965034965
```

Figure 3 shows the Confusion matrices and testing accuracies for all models

4. Conclusion

This project helped us gain insight into the one of the possible roles of machine learning in the field of healthcare. It gave me the opportunity to work with

various software's and libraries and produce effective results.

It was a great learning opportunity where we trained and tested various machine learning classifiers using a widely used dataset. Through this project, we learnt thatML classifiers serve as powerful mediums for the prediction and detection of various diseases and are frequently used in today's healthcare sector.

In the future, we want to keep honing our skills and gain more knowledge about the same. We wish to participate in similar projects in the future and try to get better results. Thankyou.

References

[1] Deepa R, Kavipraba R, Pavithra G, Preethi S,Sri Rakshitha A K. Breast CancerClassification using the Supervised LearningAlgorithms. 5th InternationalConference on Intelligent Computing andControl Systems (ICICCS), 2021.

[2] Dr.B.Santhosh Kumar, T.Daniya, Dr. J.Ajayan. Breast Cancer Prediction Using Machine Learning Algorithms. International Journal of Advanced Science and Technology, Vol. 29, No. 03, pp. 7819 -7828, 2020.

[3] Fatima, Noreen et al.Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis.*IEEE Access* 8 : 150360-150376, 2020.

[4] Gaurav Singh. Breast Cancer Prediction UsingMachine Learning. International Journal ofScientific Research in Computer Science, Engineeringand Information Technology (IJSRCSEIT), ISSN :2456-3307, Volume 6 Issue 4, pp. 278-284,2020.

[5] Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, Thomas Noel. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Computer Science,Volume 83,Pages 1064-1069,ISSN 1877-0509, 2016.

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

[7] Mohammed Amine Naji, Sanaa El Filali,Kawtar Aarika, EL Habib Benlahmar, RachidaAit Abdelouhahid, Olivier Debauche. MachineLearning Algorithms For Breast CancerPrediction And Diagnosis, Procedia ComputerScience, 2021.

[8] Rawal, Ramik.Breast Cancer Prediction using Machine Learning.Journal of Emerging Technologies and Innovative Research (JETIR), Volume 7 Issue 5, ISSN-2349-5162, 2020.

[9] Sunny, Jean & Rane, Nikita & Kanade, Rucha & Devi, Sulochana. Breast Cancer Classification and Prediction using Machine Learning. International Journal of Engineering Research and. V9. 10.17577/IJERTV9IS020280, 2020.

[10]Y.-S. Sun, Z. Zhao, Z.-N. Yang, F. Xu, H.-J. Lu, Z.-Y. Zhu, W. Shi, J. Jiang, P.-P. Yao, and H.-P. Zhu. Risk factors and preventions of breast cancer. Int. J. Biol. Sci., vol. 13, no. 11, p. 1387, 2017.