WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Cancer Detection Using Machine Learning

Kumud Saxena, Krishna Gupta, Pankaj Prajapati, Parth Shakoley, Jeewan Karki

Noida Institute of Engineering and Technology

ABSTRACT

Each year, the number of deaths is increasing tremendously due to breast cancer. It is a common type of all cancers and the major cause of death in women worldwide. Any development in predicting and diagnosing cancer disease is capital important for a healthy life. Consequently, high accuracy in cancer prediction is important to update the treatment aspect and the survivability standard of patients. Machine learning techniques can give a great contribution on the process of prediction and early diagnosis of breast cancer and have proved to be a strong technique, making it a research hotspot. In this study, we applied five machine learning algorithms: Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision tree (C4.5), and K-Nearest Neighbors (KNN) on the Breast Cancer Wisconsin Diagnostic dataset. Obtaining the results, a performance evaluation and comparison are carried out between these different classifiers. The main objective of this research paper is to predict and diagnose breast cancer using machine-learning algorithms and find out which algorithm is effective with regard to the confusion matrix, accuracy, and precision. Observations show that the Support Vector Machine outperformed all other classifiers and achieved the highest accuracy of 97.2%. All the work is done in the Anaconda environment based on python programming language and Scikit-learn library.

Introduction

According to the statistics published by IARC in December 2020, breast cancer has surpassed lung cancer to become the most commonly diagnosed cancer in women worldwide. In the last twenty years, the number of people diagnosed with cancer has nearly doubled: it has passed from 10 million in 2000 to 19.3 million in 2020. Today, one in 5 people worldwide will develop cancer during their lifetime. Projections suggest that the number of people being diagnosed with cancer will increase still further in the coming years and will be nearly 50% higher in 2040 than in 2020. The number of deaths from cancer has also increased: from 6.2 million in 2000 to 10 million in 2020. More than one in six deaths is due to cancer. This reinforces the need to invest in both the fight against cancer and cancer prevention. The ICTs' successful introduction in medical practice is an important stake in the renovation of the health system, and more precisely, in cancer care. Actually, Big data has made a big change in BI by analyzing a large amount of unstructured, heterogeneous, non-standard, and incomplete healthcare data. Moreover, it does not only predict but also helps in decision-making and is increasingly noticed as a breakthrough in ongoing advancement with the goal to improve the quality of patient care and reduces the cost of healthcare. Applied data mining algorithms play a significant role in the healthcare industry due to their high performance in predicting, diagnosing diseases, reducing costs of medicine, making real-time decisions to save people's lives. The most common data mining modeling goals are classification and prediction, which use a number of algorithms for the prediction of breast cancer. The paper, therefore, gives a comparison of the performance of five main classifiers: Support Vector Machine (SVM)), Random Forest, Logistic Regression, Decision tree (C4.5), and K-Nearest Neighbors (KNN Network), which are, according to the research community, among the most influential algorithms in data mining and in the top 10 algorithms of data mining. Our goal is to predict and diagnose breast cancer using machine-learning algorithms, finding out what is the best one based on the performance of each classifier, using criteria such as the confusion matrix, accuracy, precision, and sensitivity. The rest of this paper is organized as follows. Section 2 introduces methods and results of previous research on the diagnosis of breast cancer. Section 3 describes the methodology proposed for our work. Section 4 presents and explains in detail the experiments' results. Section 5 concludes the paper.

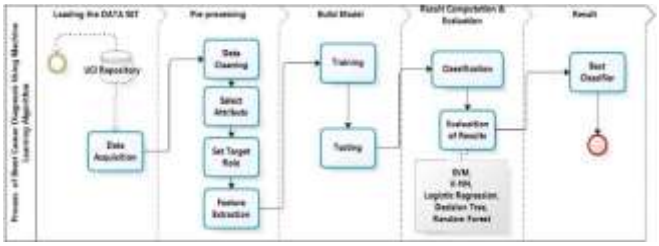
Related Works

A large number of machine learning algorithms are available for prediction and diagnosis of breast cancer. Some of the machine learning algorithms are Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree (C4.5), and K-Nearest Neighbors (KNN Network). A lot of researcher have realized research in breast cancer by using several datasets such as using SEER dataset, Mammogram images as dataset, Wisconsin Dataset, and also dataset from various hospitals. By exploiting these datasets, authors extract and select various features and complete their research. These are some significant research. The author Sudarshan Nayak demonstrates the use of various supervised machine learning algorithms in classification of breast cancer from using 3D images and finds that SVM is the best based on his overall performance. On the other side, we find that B.M. Gayathri works on comparative study of Relevance vector machine which provides Low computational cost while comparing with other machine learning techniques which are used for breast cancer detection, and explain how RVM is better than other machine learning algorithms for diagnosing breast

cancer even the variables are reduced and achieved 97% accuracy. Hiba Asri demonstrated that Support Vector Machine proves its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of precision and low error rate with an accuracy of 97.13%. In recent works, we find that Youness Khoudfi and Mohamed Bahaj, similarly proposed a comparison between Machine learning algorithms and they found the best classifier SVM with an accuracy of 97.9% compared with K-NN, RF, and NB, they are based on Multilayer Perception with 5 layers and 10 times cross-validation using MLP. The author Latchoumiet TP, found a classification value of 98.4% proposing an optimization weighting of the particle swarm (WPSO) based on the SSVM for the classification. Ahmed Hamza Osman proposed a solution for the diagnosis of WBCD with a prediction of 99.10% found by the SVM algorithm by combining a clustering algorithm with an efficient probabilistic vector support machine. Our research targets assessing such machine learning algorithms and Approaches in order to conclude the best methodology for breast cancer prediction and diagnosis.

Methodology

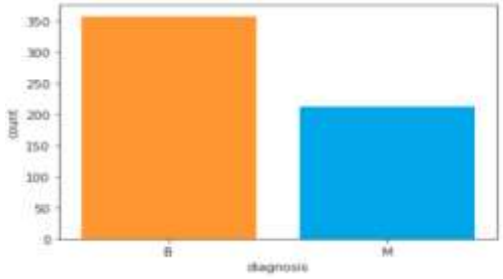
The main aim of the experiment is to find a much efficient and predictive algorithm for the detection of breast cancer; hence, we applied machine learning classifiers like Support Vector Machine, Random Forests, Logistic Regression, Decision tree, and K-Nearest Neighbors on the dataset of Breast Cancer Wisconsin Diagnostic and evaluate results obtained for defining which model provides higher accuracy.



Our methodology begins with data acquisition followed by preprocessing, which contains four steps viz: data cleaning, select attributes, set target role and features extraction. Using the prepared data, we build machine learning algorithms that will predict if there is breast cancer for a new set of measurements. To evaluate the algorithm's performance, we show the model new data for which we have labels. This is normally done by splitting the labeled data we have collected into two parts using the train_test_split method. 75% of the data is used for building our machine learning model and is normally called the training data or training set. 25% of the data will be used to access how good the model works and is called test data, test set. After testing the models, we compare the obtained results to select the algorithm, which provides the high accuracy, and identify the most predictive algorithm for the detection of breast cancer.

Machine Learning Algorithms

The predictive analysis of the machine learning algorithms is achieved in our project. The machine learning algorithms applied in our project are:



• Support Vector Machine (SVM) is a classifier which divides the datasets into classes to find a maximum marginal hyper plane (MMH) via the nearest data points .

• Random forests or random decision forests are an ensemble method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set

• k-Nearest Neighbors (K-NN) is a supervised classification algorithm. It takes a bunch of labeled points and uses them to learn how to label other points. To label a new point, it looks at the labeled points closest to that new point, which is its nearest neighbors, and has those neighbors vote [10].

• Logistic regression is a very powerful modeling tool, is a generalization of linear regression [11]. Logistic Regression is used to assess the likelihood of a disease or health condition as a function of a risk factor (and covariates). Both simple and multiple logistic regression, assess the association between independent variable(s) (Xi) -- sometimes called exposure or predictor variables — and a dichotomous dependent variable

(Y) -- sometimes called the outcome or response variable. It is used primarily for predicting binary or multiclass dependent variables.

• Decision Tree C4.5 is a predictive modeling tool that can be applied across many areas. It can be constructed by an algorithmic approach that can split the dataset in different ways based on different conditions [12].

Dataset acquisition

In our study, we use Breast Cancer Wisconsin Diagnostic dataset from University of Wisconsin Hospitals Madison Breast Cancer Database [13]. The features of dataset are computed from a digitized image of a breast cancer sample obtained from fine-needle aspirate (FNA). The characteristics of the cell nuclei present in the image are determined from these features. Breast Cancer Wisconsin Diagnostic has 569 instances (Benign: 357 Malignant: 212), 2 classes (62.74% benign and 37.26% malignant), and 11 integer-valued attributes (-Id -Diagnosis -Radius - Texture -Area -Perimeter -Smoothness -Compactness -Concavity -Concave points -Symmetry -Fractal dimension).

Experiment Environment

All the experiments performed on the machine learning algorithms described in this paper were carried out using the Scikit-learn library and Python programming language. Scikit-learn, also known as sklearn, is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

Results And Discussion

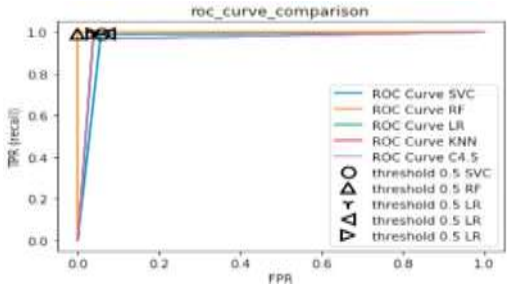
After applying Machine Learning Algorithms on Breast Cancer Wisconsin Diagnostic dataset. We used the Confusion Matrix, Accuracy, Precision, Sensitivity, F1 Score, AUC as performance metrics to evaluate and compare models to identify the best algorithm for the breast cancer Prediction. The confusion matrix is the way to measure the performance of a classification problem where output can be of two or more type of classes. A confusion matrix is a table that is used to determine the performance of a classification problem where the output can be of two or more types of classes. It's called a two-dimensional table with "Actual" and "Predicted" and further have "True Positives (TP)", "True Negatives (TN)", "False Positives (FP)", and "False Negatives (FN)". Accuracy is the most common performance metric for classification algorithms. It may be defined as the number of correct predictions made as a ratio of all the predictions made. Precision is used in document retrievals and may be defined as the number of correct documents returned by our ML model. Sensitivity may be stated as the number of positives returned by your ML model. F1 score provides us with the harmonic mean of precision and Sensitivity. The formula of F1 score is a weighted average of the precision and Sensitivity.

Table 1 and figure 2 show the accuracy percentage for Wincson Breast Cancer Diagnostic datasets. From the results in the training set and the test set, we can see that all the classifiers present different values of accuracy; however, SVM always has a higher value of accuracy in the test set (97.2%) than the other classifiers.

Table 1. Accuracy percentage for breast cancer diagnostic dataset.

Algorithms	Accuracy Training Set (%)	Accuracy Testing Set (%)
SVM	98.4%	97.2%
Radom Forest	99.8%	96.5%
Logistic Regression	95.5%	95.8%
Decision Tree K-NN	98.8%	95.1%
	94.6%	93.7%

Because the Confusion Matrices are one of the good ways toevaluate the classifier, each row of Table 2 represents the rates in an actual class while each column shows the predictions. Table 3 presents the calculated performance measures of classification models based on confusion matrix results, precision sensitivity f1 score for benign and malignant.



Conclusion

On the Wisconsin Breast Cancer Diagnostic dataset, we apply five major algorithms: SVM, Random Forests, Logistic Regression, Decision Tree, and K-NN. We calculate, compare, and evaluate the different results obtained by the confusion matrix, accuracy, sensitivity, precision, and AUC to determine the best machine learning algorithm that is accurate, reliable, and holds higher accuracy. All algorithms have been developed in Python using the scikit-learn library in the Anaconda environment. After the precise comparison between our models, we found that the Support Vector Machine achieved a higher efficiency of 97.2%, precision of 97.5%, AUC of 96.6%, and outperformed all other algorithms. Conclusion: The Support Vector Machine has shown its efficiency in breast cancer prediction and diagnosis, and achieves the best performance in terms of accuracy and precision. Note that all results obtained are related only to the WBCD database; hence, this can be considered a limitation of our work. Thus, it is necessary to reflect for future work on applying these same algorithms and methods on other databases for the confirmation of results obtained via this database. Moreover, in our future work, we plan to apply our and other machine learning algorithms using new parameters on larger data sets with more disease classes to obtain higher accuracy.

References

[1] 'WHO Breast cancer', *WHO*. http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/ (accessed Feb. 18, 2020).

[2] Datafloq - Top 10 Data Mining Algorithms, Demystified. https://datafloq.com/read/top-10-data-mining-algorithmsdemystified/1144. Accessed December 29, 2015.

[3] S. Nayak and D. Gope, "Comparison of supervised learning algorithms for RF-based breast cancer detection," 2017 Computing and Electromagnetics International Workshop (CEM), Barcelona, 2017, pp.

[4] B.M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016, pp. 1-5.

[5] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, 'Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis', *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.

[6] Y. khoudfi and M. Bahaj, Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification, 978-1-5386- 4225-2/18/$31.00 ©2018 IEEE.

[7] L. Latchoumi, T. P., & Parthiban, "Abnormality detection using weighed particle swarm optimization and smooth support vector machine," Biomed. Res., vol. 28, no. 11, pp. 4749–4751, 2017.

[8] A. H. Osman, "An Enhanced Breast Cancer Diagnosis Scheme based on Two-Step-SVM Technique," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 4, pp. 158–165, 2017.

[9] Noble WS. What is a support vector machine? Nat Biotechnol. 2006;24(12):1565-1567. doi:10.1038/nbt1206-1565.

[10] Larose DT. Discovering Knowledge in Data. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2004.

[11] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York, NY: Springer-Verlag;2001.

[12] Quinlan JR. C4.5: Programs for Machine Learning.; 2014:302. https://books.google.com/books?hl=fr&lr=&id=b3ujBQAAQBAJ&pgis=1.

[13] "UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set."

[14] Fabian Pedregosa and all (2011). "Scikit-learn: Machine Learning in Python". Journal of Machine Learning Research. 12: 2825–2830.