IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Machine Learning Algorithm for Malignant and Benign Breast Cancer Classification

Yasir Iftekhar Khan

Indian Institute of Technology Mandi, Himachal Pradesh.

Abstract- Currently the most common form of cancer diagnosed in women all over the world is breast cancer. One of the leading causes of death for women, it arises in breast tissue. If this cancer is discovered in its early stages, it can be treated. There are two types of tumors that can occur in breast cancer patients: malignant and benign. Malignant tumors are deadly because its growth rate is much higher than the benign one. Therefore, early identification of tumor type is essential for adequate treatment of a patient with breast cancer. In this work, the Wisconsin breast cancer data set was used, which was collected from the UCI repository. The goal is to analyze the data set and evaluate the performance of various machine learning Breast cancer prediction algorithms, here support vector Machine, logistic regression, K-nearest neighbors, Gradient Boosting Classifier and Random Forest classifiers have been implemented to classify tumors into benign and malignant. To determine the most appropriate algorithm, the accuracy of each is calculated and compared. Accordingly, Logistic Regression classifier has the highest accuracy of 99.12%. These classifiers can be used to build an automatic diagnostic system for preliminary breast diagnosis cancer.

Keywords- Breast Cancer diagnosis, Analysis of WBCD Dataset, Malignant-Benign breast cancer classification, Machine Learning Algorithms.

I. INTRODUCTION:

Breast cancer is a type of cancer that develops in the breast (BC). Cancer is brought on by unchecked cell division or growth. Breast cancer cells typically form a tumor that can be seen on an X-ray. One of the most common diseases that cause death in women is breast cancer. Breast cancer, one of the most frequent cancers in women, consistently exhibits a very high incidence and death rate, impacting 10% of women at various phases of their life. It is the second leading cause of death for women after lung cancer. Among all cancers, breast cancer accounts for 25% overall with 12% of all new cases in women [1]. Breast cancer can be detected by classifying tumors. As seen in breast cancer cases, there are two types of tumors: malignant and benign. Malignant tumors spread faster than benign tumors. Doctors require a reliable diagnostic technique to distinguish between these tumors. However, even specialists have difficulty distinguishing tumors in most cases. As a result, a dependable automatic diagnostic system is critical for tumor type diagnosis. To determine breast cancer, patients are frequently subjected to an avalanche of tests including ultrasound, biopsy and mammography depending on the variable nature of the breast symptoms of cancer. Among these methods, the most indicative is biopsy involving the removal of samples of tissue or cells for research. The sample of cells is taken from a breast by the fine-needle aspiration (FNA) procedure and then sent for analysis under a microscope to a pathology laboratory [2]. Numerical characteristics such as radius, texture, perimeter and area can be calculated from microscope images of cells and tissues. Data obtained from FNA are then analyzed in conjunction with various imaging data to predict the likelihood of the patient having a malignant tumor. Early diagnosis of breast cancer can significantly improve the prognosis and probability of survival because it allows patients to receive timely clinical treatment. The correct identification of breast cancer and the categorization of patients into malignant or benign classes is a very important line of research. Various methods to predict breast cancer have been established in recent years. Classification techniques, for example, Random Forest (RF), Support Vector Machine (SVM), Adaboost Classifier and K Nearest Neighbors (KNN have been used in recent literature

The Wisconsin Breast Cancer Dataset (WBCD) of the FNA biopsy system was used in this study, and various machine learning (ML) classifiers were implemented to determine the type of breast cancer in a suspected patient. Six classification technique is used including Random Forest, Logistic Regression, Gradient Boosting, XGBClassifier, SVM and KNN. The obtained results are then evaluated in order to compare the algorithms in order to find the best model for predicting breast cancer. The main objective of our article includes analyzing the WBCD dataset to discover relationships between features, comparing different established classifiers on the WBCD dataset, and determining the most satisfactory approach supporting the dataset with high prediction accuracy. The rest of the paper includes Literature review, overall methodology, results obtained and conclusion.

II. LITERATURE REVIEW:

Many models have been proposed in previous works. that use different feature sets and machine methods Learn to diagnose breast cancer. The rarity of the great data sets and inequality between negative and positive classes are the main challenges in the field of breast cancer research prediction. In [1], the main objective of the analysis is to detect the algorithm that works more quickly, reliably and efficiently to predict breast cancer. With an accuracy of 99.76%, Random Forest outperforms all other algorithms. In [2], [4], [5], authors carried out a comparative analysis of breast cancer prediction offered by existing ML algorithms. The data set used in these articles is WBCD. The authors of [6] used a Random Forest classifier to identify and predict breast cancer in order to determine whether or not a person has breast cancer. Because the Random Forest algorithm employs both classification and regression approaches, it provides the highest level of identification accuracy. The authors of [7] provided a study on breast cancer to develop predictive models for breast cancer survival. In this study, three breast cancer survivability prediction models were applied to two types of cancer: benign and malignant. The authors of [8] summarized all previous research on ML algorithms used for breast cancer detection. They proposed that data augmentation techniques could solve the problem of limited available datasets. The authors of [9] presented a technique for detecting and identifying cell morphology in automated systems that perform classification using computer-aided mammogram image features. In [10], the authors compared Various sorting and clustering algorithms in the survey. The result shows that the classification algorithms are better predictors than clustering algorithms. In [11], the automatic anomaly detection method in mammograms is discussed. Apply the middle blur C and the thresholding strategy, the suspected region of interest (ROI) was segmental. The proposed algorithm for the Mini-MIAS dataset has been validated. They concluded that the detection performance of suspicious regions in mammograms can be improved by subtracting the enhanced preprocessing and then enhanced preprocessing reverse images. In [12], for the identification of malignant tumor and benign state, the authors have proposed an algorithm according to a fuzzy inference system. Comparison of classical performance criteria such as sensitivity, precision and specificity suggest that your introduced solution outperforms Artificial neural network (ANN) and SVM classification. In [13], different deep learning concepts related within Analyzes of mammograms and contributions to this article are reviewed. domain is summarized. This book summarizes the past of mammography research, recent advances and current status art.

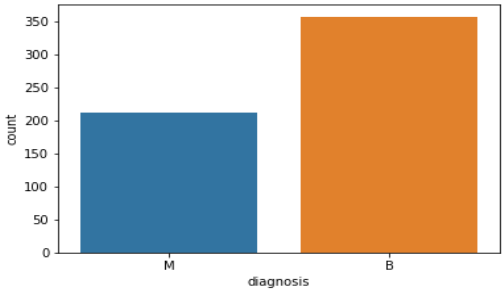
III. METHODOLOGY:

A series of step is set up to get to the most reliable results to determine if the stage of the tumor is malignant (cancerous) or benign (not cancerous). Our general methodology can be presented below subsections: -

- Dataset Description
- Dataset Analysis
- Training and Testing

A. Data Description:

Dr. William H. Wolberg of the University Hospital of Wisconsin in Madison, Wisconsin, USA, developed the WBCD dataset used for this article that is publicly available. East the data set includes 357 and 212 cases of benign and malignant diseases breast cancer respectively as shown in the below Fig.



The data set includes 32 columns, with the identification number being the first column and the result of the diagnosis (0-benign and 1-malignant) being the second column. The rest of columns (3-32) contain three measurements (average, standard deviation and mean of the worst) of ten characteristics. These characteristics they represent the shape and size of the nucleus

of the target cancer cell. Cell sample is taken from a breast by Fine Needle aspiration procedure (FNA) in biopsy test. For each cell nucleus, these characteristics are determined by analyzing under a microscope in a pathology lab. All the values of the characteristics are stored up to four significant digits. There was there are no null entries in the data set. The 10 characteristics

Feature Name	**Feature Description**
Radius	Mean of distances from center to points on the perimeter
Texture	Standard deviation of gray-scale values
Perimeter	Gross distance between the snake points
Area	Total number of pixels on the inside of the snake along with one half of the pixels in the circumference
Smoothness	Local variation in radius lengths
Compactness	Perimeter^2 / area - 1.0
Concavity	Severity of concave portions of the contour
Concave Points	Number of concave portions of the contour
Symmetry	The difference in length between line perpendicular to the major axis in both directions to the cell boundary.
Fractal dimension	Coastline estimation

are described in the below table

B. Data Analysis:

For dataset analysis, the entire dataset was considered. Now in Fig. 1, the correlation between the features of the WBCD dataset is displayed in a heat map. Correlation heat map displays a 2D correlation matrix between two discrete dimensions where the value of the first dimension is considered as a line and the value of the second dimension as a column of the heat map. The dimensional value (-1 to +1) is calculated from the linearity between the pair of features. If the two variables vary and evolve in the same direction, positive correlation is acquired. In case of negative correlation, increase by one variable is associated with a decrease in the other and vice-versa. In Fig. 2 correlation with the diagnosis column is shown.

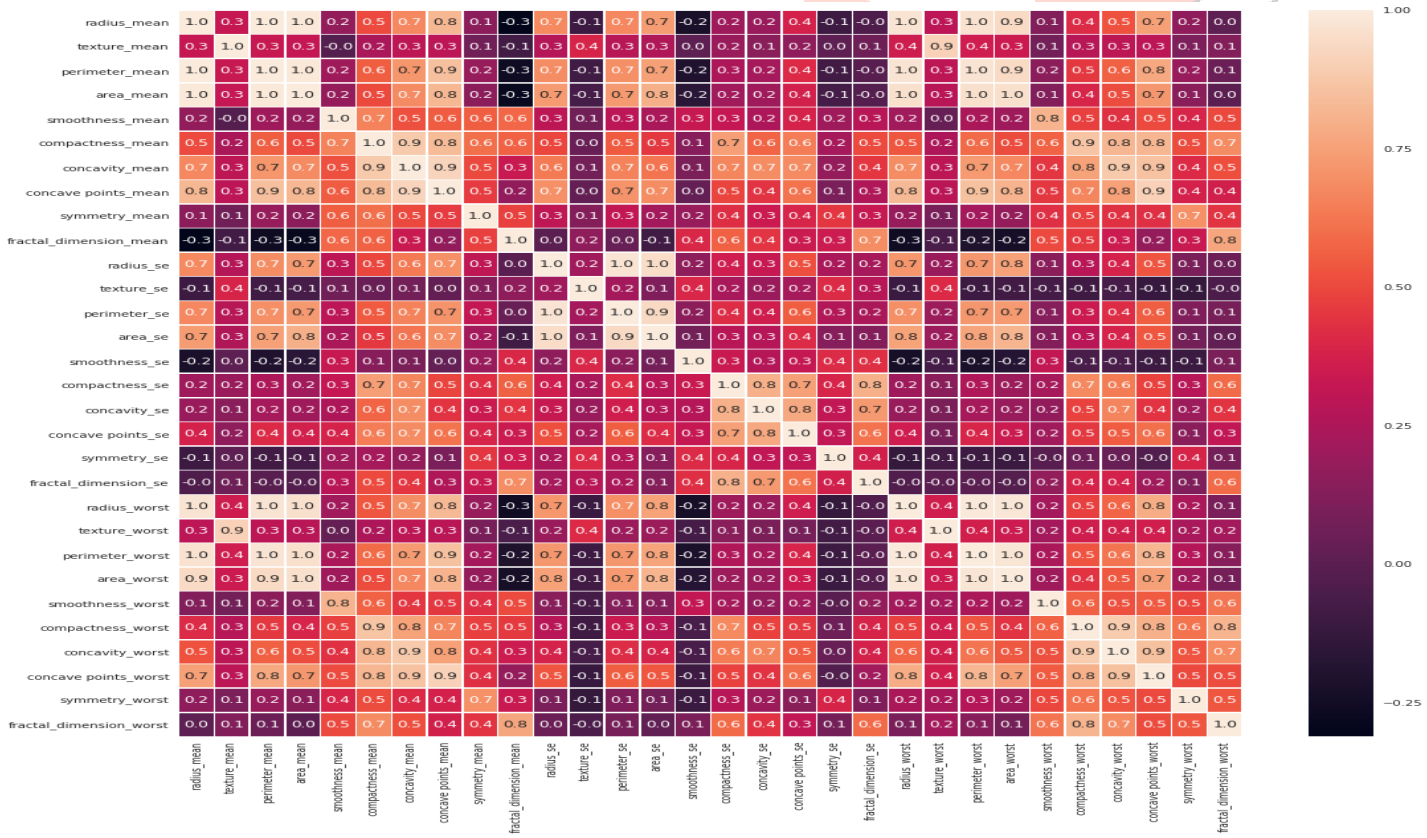


Figure 1

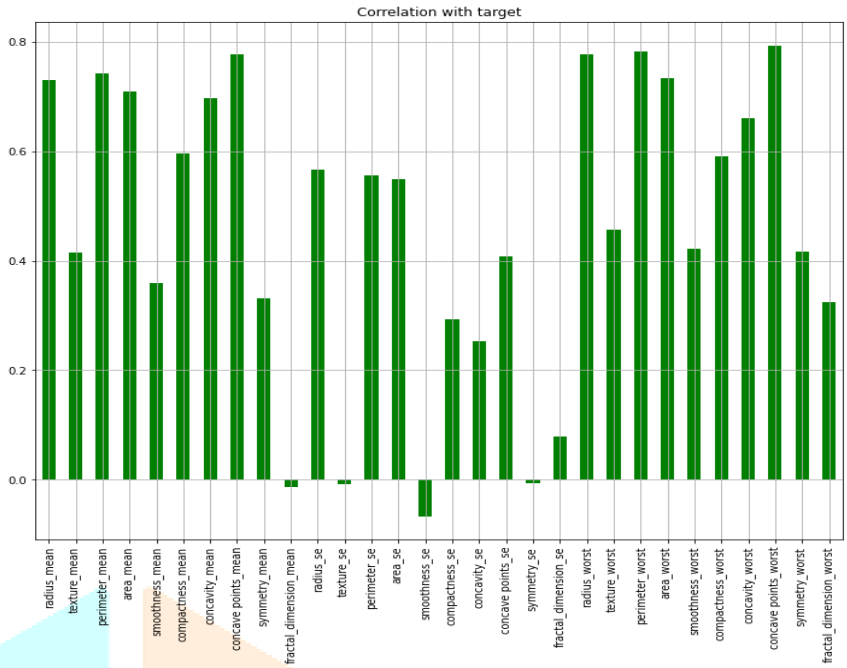


Figure 2

C. Training and Testing:

Initially, the dataset is read from the CSV file. The data the dataset entries are parsed based on their features before they are used for a later step. then we parted ways the two-part random dataset: training set (80%) and test equipment (20%). For training the model either the column with the greatest correlation to the target can be use, or the whole columns can be use. But for this experiment all the column is used. But before training the model, the outliers have to be removed. This is done using Local Outlier Factor (LOF). The LOF algorithm is an unsupervised anomaly detection method that calculates the local density deviation of a given data point from its neighbors. Consider as outliers those samples that have a significantly lower density than their neighbors. For this experiment the ML models used are Logistic Regression, K-Neighbor Classifier (KNN), Support Vector Machine (SVM), Random Forest Classifier, Gradient Boosting Classifier, XGBClassifier.

IV. RESULTS:

In this section, after implementing the ML algorithms, we have analyzed the performance of the algorithms on the data set. This is achieved by running the algorithms on the old define test data set. 20% of all data was included in test data set. A confusion matrix is generated for the actual and expected result composed of TP (True Positive), FP (False Positive), TN (True Negative) and FN (False Negative) for accuracy calculation for each algorithm used.

The formula for the accuracy is:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN).$$

The below shown table demonstrate the accuracy obtained for different ML models used in the experiment

Algorithms	Accuracy Obtained
Logistic Regression	99.12%
KNN	95.37%
SVM	96.50%
Random Forest Classifier	96.20%
Gradient Boosting Classifier	97.60%
XGBClassifier	97%

Hence, we can conclude that Logistic Regression performs the best and gives the accuracy of 99.12% while the KNN gives the least accuracy of 95.37%. Hence for this experiment Logistic regression is the preferred choice.

V. CONCLUSION:

One of the deadly diseases affecting women these days is breast cancer. The Wisconsin Breast Cancer Dataset was used in our study, and several ML algorithms were used to assess the efficacy and usefulness of these algorithms in classifying malignant and benign breast cancer. The correlation between the different characteristics of the dataset was analyzed for feature selection. The results will help to choose the best ML algorithm to build of an automatic system for diagnosing breast cancer. Of our study, we can conclude that Logistic Regression gives Maximum accuracy with 99.12% accuracy. In the future, a way to improve this work is by managing a fairly large dataset and including more features like breast cancer phase identification.

REFERENCES:

1. J. Sivapriya, A. Kumar, S. Siddarth Sai, and S. Sriram, "Breast cancer prediction using machine learning," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, 2019.
2. Y. Khourdifi and M. Bahaj, "Applying best machine learning algorithms for breast cancer prediction and classification," in 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS). IEEE, 2018, pp. 1–5.
3. N. K. Sinha, M. Khulal, M. Gurung, and A. Lal, "Developing a web based system for breast cancer prediction using xgboost classifier," International Journal of Engineering Research Technology (IJERT), vol. 9, 2020.
4. R. Dhanya, I. R. Paul, S. S. Akula, M. Sivakumar, and J. J. Nair, "A comparative study for breast cancer prediction using machine learning and feature selection," in 2019 International Conference on Intelligent Computing and Control Systems (ICCS). IEEE, 2019, pp. 1049–1055.
5. M. M. Islam, H. Iqbal, M. R. Haque, and M. K. Hasan, "Prediction of breast cancer using support vector machine and k-nearest neighbors," in 2017 IEEE Region 10 Humanitarian Technology Conference (R10- HTC). IEEE, 2017, pp. 226–229.
6. M. S. Yarabarla, L. K. Ravi, and A. Sivasangari, "Breast cancer prediction via machine learning," in 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, 2019, pp. 121–124.
7. [V. Chaurasia, S. Pal, and B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," Journal of Algorithms & Computational Technology, vol. 12, no. 2, pp. 119–126, 2018.
8. N. Fatima, L. Liu, S. Hong, and H. Ahmed, "Prediction of breast cancer, comparative review of machine learning techniques, and their analysis," IEEE Access, vol. 8, pp. 150 360–150 376, 2020.
9. A. Toprak, "Extreme learning machine (elm)-based classification of benign and malignant cells in breast cancer," Medical science monitor: international medical journal of experimental and clinical research, vol. 24, p. 6537, 2018.
10. D. S. Jacob, R. Viswan, V. Manju, L. PadmaSuresh, and S. Raj, "A survey on breast cancer prediction using data miningtechniques," in 2018 Conference on Emerging Devices and Smart Systems (ICEDSS). IEEE, 2018, pp. 256–258.
11. K. L. Kashyap, M. K. Bajpai, and P. Khanna, "Breast cancer detection in digital mammograms," in 2015 IEEE international conference on imaging systems and techniques (IST). IEEE, 2015, pp. 1–6.
12. F.-T. Johra and M. M. H. Shuvo, "Detection of breast cancer from histopathology image and classifying benign and malignant state using fuzzy logic," in 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT). IEEE, 2016, pp. 1–5.
13. O. V. Singh and P. Choudhary, "A study on convolution neural network for breast cancer detection," in 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP). IEEE, 2019, pp. 1–7.