International Journal of Research in Engineering, Science and Management
Volume 6, Issue 4, April 2023
https://www.ijresm.com | ISSN (Online): 2581-5792

Breast Cancer Detection using Machine Learning Classifier Algorithms

Ravi Hemraj Gedam[1], Syed Mohammed Saflan Ali[2*], Vivian Rodrick James[3],
Mrunal Rajendra Sonekar[4], Mayur Namdev Choudhary[5]
[1]Assistant Controller of Examination, Department of Computer Science and Engineering, G.H. Raisoni University, Chhindwara, India
[2,3,4,5]Student, Department of Computer Science and Engineering, G.H. Raisoni University, Chhindwara, India

***Abstract*: Breast cancer is a prevalent form of cancer among women globally, particularly in developing nations where most diagnoses occur in the later stages of the disease. It is one of the most dangerous types of cancer that affects women. Cancer.net offers personalized pathways for over 120 types of cancer and genetic diseases. Previous projects have compared machine learning algorithms using various techniques such as ensemble methods, data mining algorithms, or blood analysis. This paper aims to compare six machine learning algorithms, namely Naive Bayes, Random Forest, Artificial Neural Networks, Nearest Neighbor, Support Vector Machine, and Decision Tree on the Wisconsin Diagnostic Breast Cancer dataset (WDBC) extracted from the cancer RAW CSV provided by Indian AI Productions. The dataset was divided into a training and testing phase to implement the ML algorithms. The algorithm that produces the best results will be used to classify cancerous tumors as benign or malignant based on their shape, size, texture, and smoothness.**

***Keywords*: breast cancer, data mining, machine learning, neural networks, WBCD, blood analysis.**

1. Introduction

Breast cancer is a prevalent form of cancer affecting women globally, as reported by the World Health Organization (WHO). It is a leading cause of death among women, with a high mortality rate in India at approximately 14%, making it the most common cancer among Indian women. Although it only affects 5% of women in India, it affects a significantly higher percentage of women in Europe and the United States at 12.5%. However, women in Malaysia tend to present with breast cancer at a later stage of the disease compared to women in other countries. Early diagnosis of breast cancer is crucial, as it is easily diagnosable when symptoms appear.

Machine learning (ML) is a subfield of artificial intelligence (AI) and computer science that utilizes algorithms and data to simulate human learning and improve accuracy gradually. In the context of breast cancer diagnosis, ML classifiers collect datasets and apply various ML algorithms to construct the foundation of the project, leading to increased accuracy and efficiency. This approach aims to reduce the traditional screening time required.

ML can also help with early diagnosis of breast cancer by analyzing data such as tumor size, shape, and color retrieved from datasets, thus determining the nature of the cancer. Studies show that female breast cancer patients who begin treatment within 90 days after symptom onset.

Have a significantly higher chance of survival than those who delay treatment. Early detection and timely treatment increase the likelihood of survival and prevent cancerous cells from spreading throughout the body.

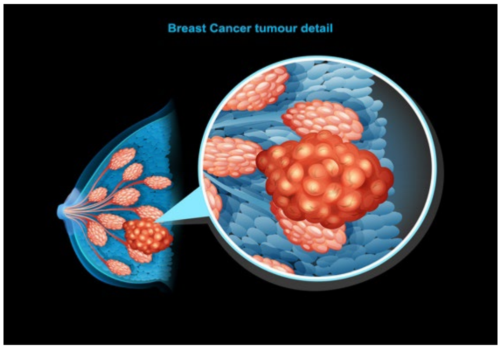


Fig. 1. Depicts the side view of the tumor detail

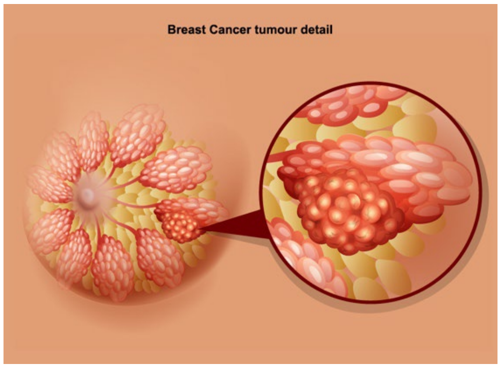


Fig. 2. Depicts the front view of the tumor detail

Accurately predicting breast cancer is still a challenging task as it requires a model that accounts for all the known risk factors. The current prediction models may only concentrate on analyzing mammographic images or demographic risk factors while ignoring other essential factors. This could lead to multiple screenings and invasive sampling procedures like magnetic resonance imaging (MRI) and ultrasound. Such

*Corresponding author: saflan0423@gmail.com practices may cause financial and psychological strain on patients.

2. Methodologies

A five-phase methodology has been developed to enhance Breast Cancer Detection recognition. The methodology comprises the following phases: gathering the Cancer dataset, analyzing features, and cleaning data, imputing missing values, normalizing data, and selecting features, and training the classification model on a training set and assessing the model on a test set. The model is evaluated on the test set to ensure its efficacy. Figure 3 illustrates the adopted methodology.

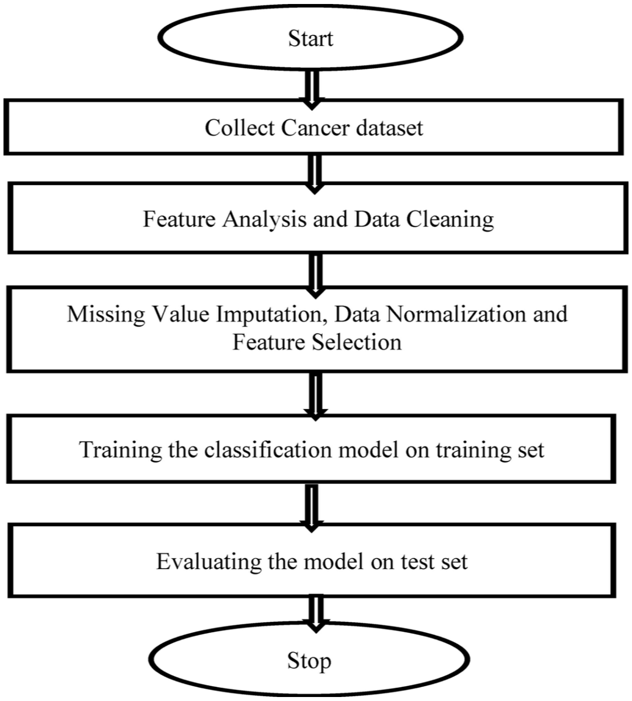

Start
Collect Cancer dataset
Feature Analysis and Data Cleaning
Missing Value Imputation, Data Normalization and Feature Selection
Training the classification model on training set
Evaluating the model on test set
Stop

Fig. 3. Methodology used for breast cancer detection

1) Start

We did our project by analyzing & studying in detail about breast cancer.

2) Collecting cancer dataset

The dataset used for our research was obtained from Indian AI Productions. We selected this dataset because of its standardization and content richness, which allowed us to compare our results with existing research. In addition to this dataset, we also referred to multiple other datasets. As we continue our research, we may expand our analysis by including data from forthcoming years.

3) Feature analysis and data cleaning

Feature analysis is an essential step in machine learning, where relevant features are automatically selected based on the type of problem being solved. This involves analyzing individual characteristics or features of every object or pattern encountered and including or excluding important features without altering them.

Data cleaning, on the other hand, involves the preparation of data for analysis by removing irrelevant or incorrect information. This data can have a negative impact on the model or algorithm, potentially leading to false conclusions.

4) Missing value imputation, data normalization & feature selection

One way to manage incomplete datasets is to eliminate cases or observations that have one or more missing values, which is referred to as case deletion or listwise deletion. However, this approach is only appropriate if the quantity of missing data in the dataset is relatively small.

In contrast, missing value imputation entails using statistical or machine learning methods to estimate missing values with discrete or continuous values. This process is similar to a pattern classification task, where the objective is to predict the missing values based on the accessible data.

Data normalization is a technique used in machine learning to standardize the values of numeric columns in a dataset to a common scale during data preparation. Although not always necessary for every dataset in a model, it is crucial when the machine learning model's features have different ranges.

Feature selection, on the other hand, involves reducing the input variables to your model by utilizing only relevant data and removing noise from the data. This aids in improving the machine learning model's performance.

5) Training the classification model on training set

Once the data has been prepared through feature analysis, cleaning, imputation, normalization, and selection, the next step is to train and test the dataset with various machine learning algorithms. This is done in order to determine the best algorithm or combination of algorithms to use for the particular problem being solved.

6) Evaluating the model on test set

Eventually, the model would be evaluated using various machine learning classifiers to achieve accuracy and effectiveness based on the chosen parameters. Additionally, we would import all necessary packages as we are utilizing Jupyter Notebook.

7) Stop

3. Implementation

A. Support Vector Machine

The Support Vector Machine (SVM) is a widely used supervised Machine Learning Algorithm for addressing classification and regression problems. It is particularly useful in high-dimensional spaces and can solve both linear and non-linear problems. SVMs are also employed for detecting outliers. Our approach involves utilizing SVM classifiers to enhance the objective of attaining superior generalization performance.

B. K – Nearest Neighbor Classifier

We have selected KNN as our algorithm of choice due to its intuitive nature and ease of implementation. Our task involves predicting the type of tumor based on its size, shape, and texture, which may be categorized as Malignant or Benign. KNN's classifier is renowned for its predictive capabilities, as it can process vast amounts of data to generate prompt results.

C. Naive Bayes Classifier

This falls into the category of supervised learning algorithms that are applicable in text classification tasks. Naive Bayes is a type of probabilistic classifier that bases its predictions on the likelihood of an object. The rationale behind opting for this classifier is its established reputation for being swift, precise, and dependable as per statistical data. Naive Bayes performs remarkably well in the realm of Natural Language Processing (NLP) issues.

D. XGBoost Classifier

The Extreme Gradient Boosting algorithm is widely recognized as a potent machine learning tool for tasks involving classification and regression, owing to its remarkable speed and predictive accuracy. Its impressive performance owes much to its optimization for parallel processing on both CPU and GPU architectures.

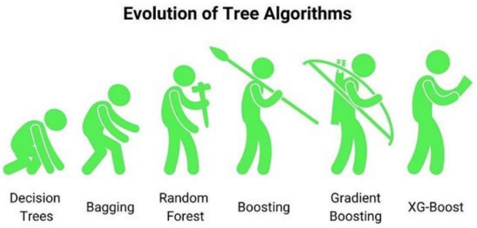


Fig. 3. Evolution of tree algorithms

The functioning of the algorithm involves a repetitive process of training several weak decision trees on the residuals of the prior trees. The primary aim is to minimize a loss function that quantifies the deviation between the predicted and actual values.

4. Results

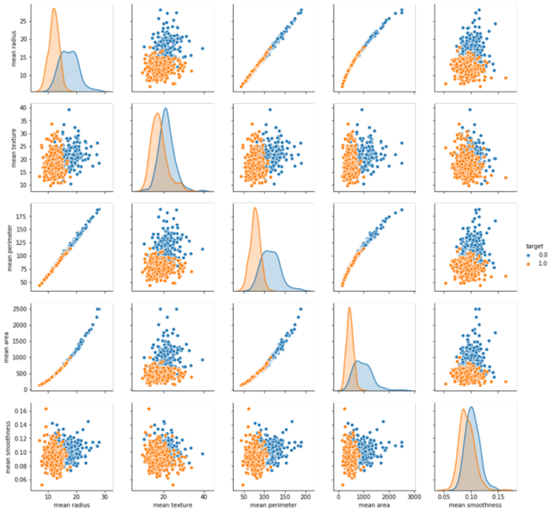
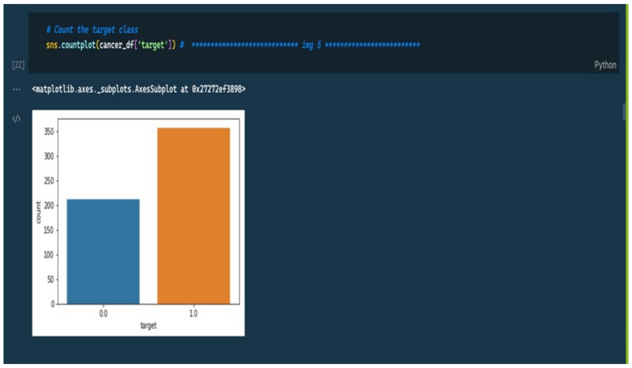


Fig. 5. Pair plot of breast cancer data

The fig. 5, displays a pair plot of breast cancer data, showcasing the distribution of malignant and benign tumor data in two distinct classes. This clear separation between the two classes in the pair plot makes it easier to differentiate between them.



```
# Count the target class
sns.countplot(cancer_df['target']) # **************************** img 5 *************************
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x27272ef3898>
```

Fig. 6. Counterplot

This fig. 6, shows the total count of malignant as well as benign tumor patients.

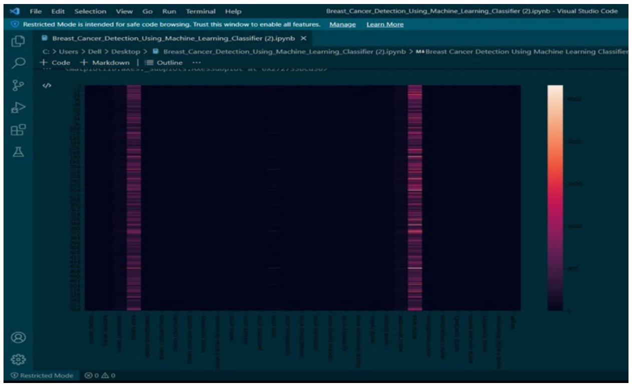
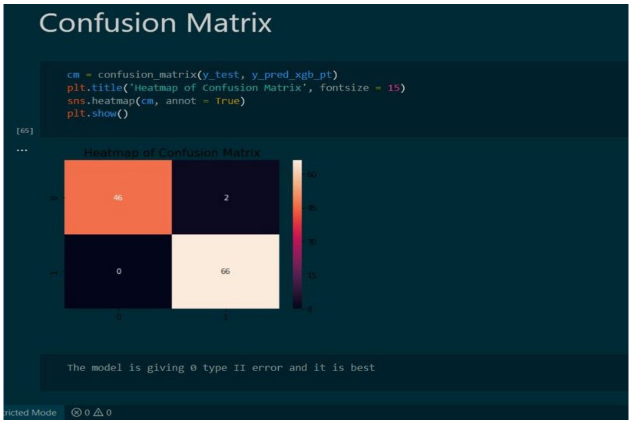

File Edit Selection View Go Run Terminal Help
Breast_Cancer_Detection_Using_Machine_Learning_Classifier (2).ipynb - Visual Studio Code
Restricted Mode is intended for safe code browsing. Trust this window to enable all features. Manage Learn More
Breast_Cancer_Detection_Using_Machine_Learning_Classifier (2).ipynb
+ Code + Markdown | Outline
Restricted Mode

Fig. 7. Heatmap of breast cancer data frame

The most vital advantage of heatmap visualization is that it sanctions, data to be presented visually which allows us to easily devour information and make more sense of it.

We can observe the variations of different feature's values as well as the feature values 'mean area' along with 'worst area' are substantial than other and 'mean perimeter', 'area error', as well as 'worst perimeter' value slightly lower but, substantial than the remaining features.



Confusion Matrix

```
cm = confusion_matrix(y_test, y_pred_xgb_pt)
plt.title('Heatmap of Confusion Matrix', fontsize = 15)
sns.heatmap(cm, annot = True)
plt.show()
```

```
The model is giving 0 type II error and it is best
```

Fig. 8. Confusion matrix

A confusion matrix is a tabular representation that is utilized to evaluate the effectiveness of a classification algorithm. This matrix provides a concise and informative overview of how well a classification model is performing. Essentially, a confusion matrix presents a summary of the predicted versus actual outcomes of a classification task. The model above is giving us 0% type II error; hence, it is the best for our use case.

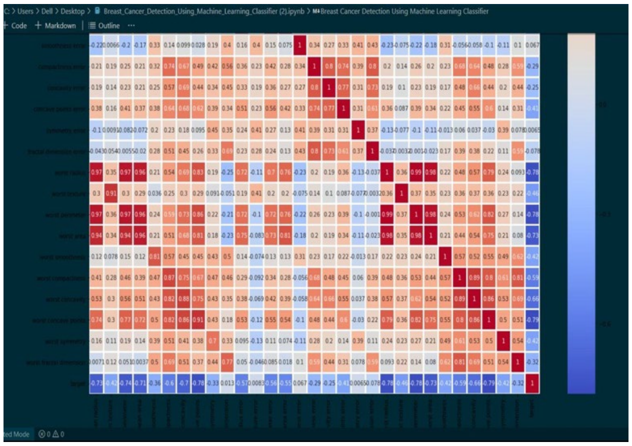

C: > Users > Dell > Desktop > Breast_Cancer_Detection_Using_Machine_Learning_Classifier (2).ipynb > Breast Cancer Detection Using Machine Learning Classifier
+ Code + Markdown | Outline ...

Fig. 9. Correlation Matrix

A correlation matrix is a structured table that displays the correlation coefficients between various variables in a dataset. This matrix outlines the correlation between all possible combinations of values within the data. The correlation matrix is a useful and efficient method for summarizing large amounts of data and for visualizing patterns and relationships within the data.

5. Conclusion

Breast cancer is a worldwide concern that has been the subject of recent studies and discussions. While developed countries such as the UK and US have seen a decrease in death rates due to advanced diagnostic technologies and awareness programs, developing countries like India continue to face challenges regarding the disease. Therefore, it is crucial to take immediate and effective measures to address this situation.

To this end, a study was conducted to explore methodologies for early detection of breast cancer utilizing a breast cancer dataset. The results indicate that various classifiers and algorithms can accurately detect and classify breast cancer data. Accurate identification of symptom properties can significantly improve detection rates. Moreover, combining multiple risk factors in modeling for breast cancer prediction can facilitate early diagnosis and the development of necessary care plans.

Efficient disease management requires the collection, storage, and management of diverse data, as well as intelligent systems based on multiple factors for predicting breast cancer. While the study utilized machine learning for breast cancer, there are numerous datasets available for different types of cancer that can be effectively detected and treated with timely diagnosis. The accuracy of detection can be further improved by applying different algorithms to the available data. Early diagnosis plays a vital role in cancer treatment, and exploring various algorithms can aid in enhancing the accuracy of detection.

References

[1] Omondiagbe, D.A. and Veeramani, S. (2019) Machine learning classification techniques for breast. IOP Science.

[2] Indian AI Production. (2019, November 20). ML project: Breast cancer detection using machine learning classifier. Indian AI Production. Retrieved from https://indianaiproduction.com/breast-cancer-detection-using-machine-learning-classifier/

[3] S;, D.M.W.L.C.P. Breast MRI for Cancer Detection and characterization: A review of evidence-based clinical applications, Academic radiology. U.S. National Library of Medicine.

[4] Stolovitzky, G. and Ben-Ari, R. (2020) The potential benefits of AI for Breast Cancer Detection, IBM Research Blog. Available at: https://www.ibm.com/blogs/research/2020/03/benefits-ai-for-breast-cancer-detection/

[5] Bhise, S. et al. (2021) Breast cancer detection using machine learning techniques, International Journal of Engineering Research & Technology.

[6] Christensen, K. (2021) Machine learning reduces uncertainty in breast cancer diagnoses, Michigan Technological University. Available at: https://www.mtu.edu/news/2021/11/machine-learning-reduces-uncertainty-in-breast-cancer-diagnoses.html

[7] Arooj, S. et al. (2022) Breast cancer detection and classification empowered with transfer learning, Frontiers in public health. Frontiers Media S.A.

[8] M. Amrane, S. Oukid, I. Gagaoua and T. Ensarİ, "Breast cancer classification using machine learning," 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, Turkey, 2018, pp. 1-4.