Breast Cancer Prediction Using Machine Learning

Detty Stanly
Department of Computer Applications
Amal Jyothi College of Engineering, Koovapally
Kottayam, Kerala
dettystanly2022@mca.ajce.in

Ms. Ankitha Philip
Asst. Professor in Computer Applications
Amal Jyothi College of Engineering, Koovapally
Kottayam, Kerala
ankithaphilip@amaljyothi.ac.in

***Abstract*-Breast Cancer is considered to be a major threat to women which leads to an increase in death rates of women that causes a major concern in the community. Even though this disease is a major threat but today medical-science is capable enough to rehabilitate such threats without causing any harm to women if detected at early stages using their innovative thoughts. Dredging the cancer and transforming between the diagnosis that certify whether the patient has breast cancer or not is considered to be the major provocation. Women worldwide are distressed by breast cancer, a widely affecting health issue that can use a large number of deaths. This paper aims to review and present an approach to identify the accuracy of breast cancer using Machine Learning. The objective is to investigate the application of multiple algorithms based on Machine Learning approach for early breast cancer detection. Machine learning algorithms are needed to identify cancers based on a given set of data, that is why automation is needed. It aims to make computers capable of self-learning. Other than relying on pre-programmed models, it is based on identifying patterns in observed data to predict outcomes.**

***Keywords* – machine learning, support vector machine, correlation, KNN algorithm, 10-fold cross validation, confusion matrix.**

I INTRODUCTION

Cancer is a disease in which some of the body's cells enlarge disorganized and replicate to other parts of the body. Cancer is the vital cause of demise in general. Bone, Lung, liver and stomach cancer are the one observed in men. While the common types seen in women are affected are breast, lung, cervical and thyroid. Breast cancer evolve as a result of genetic mutations or damage in the structure of DNA. These can be correlated with exposure to estrogen, inherited genetic defects, or genes that can cause cancer. When a person is healthy, their immune system hits any abnormal growths. When a person has cancer, this does not happen. As a result, cells within breast tissue begin to replicate uncontrollably, and they do not die as usual. This immoderate cell growth generates a tumor that dispossess nearby cells. Breast cancer generally begins off within the internal lining of the milk ducts that deliver milk.

The shortage of opinion models results in difficulty for professionals to conduct a treatment scheme that may extend patient's survival time. Hence we need time to develop a skill that can give a lower limit of error to increase accuracy. The traditional model to diagnose this cancer like mammogram and ultrasound were time- consuming. So there was a necessity for a technically feasible diagnosis in which Machine Learning methodology was implemented. This methodology includes algorithms that help for the classification of the tumor and detect the cells more precisely and take less time as well.

II. MACHINE LEARNING ALGORITHMS

Machine learning is the technology of having computer systems to behave without being clearly prearranged. Machine learning is a branch of artificial intelligence and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. It is a technique that is used to analyze data that automates analytical model building. We implement Machine Learning based on the concept that the system can study from a given data, identify them and make decision with a low human interrelation. It focuses on data and algorithms to follow the way which humans learn and try to improve the accuracy. With the furtherance brought in Machine Learning, detection and prediction of breast cancer can be done with less effort. It can be used to diagnose not only cancerous diseases but also other medical conditions. Machine Learning can be classified into three categories:

a. Supervised Learning
b. Unsupervised Learning
c. Reinforcement Learning

Supervised Learning is the most basic among the other types. Here a model is trained on labeled data. It is extremely powerful enough to use in right circumstances. It will always continue to improve even if it is deployed.

In unsupervised learning, we can study the models even without tagging the data. Since there is no labeling it results hidden structures which makes this learning distinct from others.

DOI: 10.5281/zenodo.6228080
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

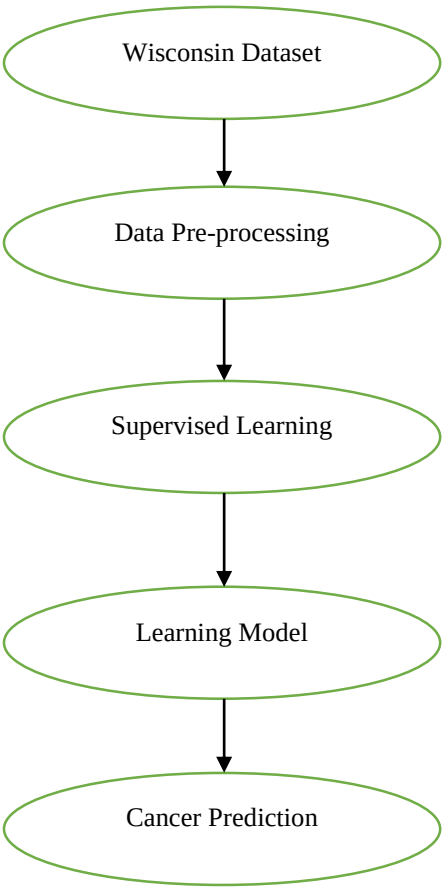
III. SUPPORT VECTOR MACHINE

Support vector machines are learning system which has many desirable qualities that make it one of the most popular algorithms. It is considered as the most accurate algorithm for classification. They are mainly based on the idea of finding hyperplanes that best divides a data set into two classes. SVM can be used to model highly complex relationships The best decision boundary is termed as the hyperplane. SVM chooses the extreme vectors that help in creating the hyperplane. Theses extreme cases are called vectors. It is used in

a. Classification
b. Pattern recognition
c. Numeric prediction

IV. PROPOSED METHODOLOGY

The purpose of this paper is to reach in a conclusion to find a Machine Learning model which can predict whether the patient is cancerous or not. The model should be able to classify correctly the dataset. The given dataset is divided into training set and testing data sets. The structure of the study is as follows:



a. Dataset

The data used for modelling is from Kaggle. The dataset is from Wisconsin Database

Table 1 : Dataset Description

Sl No.	Attributes
1	Id
2	Diagnosis
3	Radius mean
4	Texture mean
5	Perimeter mean
6	Area mean
7	Smoothness mean
8	Compactness mean
9	Concavity mean
10	Concave points mean
11	Symmetry mean
12	Fractal dimension mean
13	Radius se
14	Texture se
15	Perimeter se
16	Area se
17	Smoothness se
18	Compactness se
19	Concavity se
20	Concave points se
21	Symmetry se
22	Fractal dimension se
23	Radius worst
24	Texture worst
25	Perimeter worst
26	Area worst
27	Smoothness worst
28	Compactness worst
29	Concavity worst
30	Concave points worst
31	Symmetry worst
32	Fractal dimension worst

b. Exploratory analysis

The dataset is loaded. Exploratory analysis is an approach of analysing datasets to summarize their main characteristics. We analyse them using visualization and graphic techniques. A quick pre-processing is done in this stage along with it which is the cleaning of data. It is the most vital process in the whole Machine Learning Methodology. The missing data and the impurities in the dataset can be reduce the quality of the output that can be performed in order to improve the effectiveness. First issue was to itemize the analysis column.

DOI: 10.5281/zenodo.6228080
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

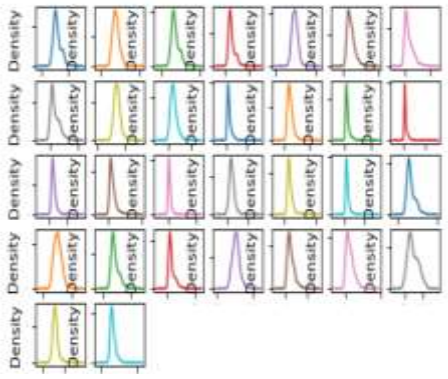


Figure 1: Distribution

It is always effectual to observe the relation between the features. Correlation methodically depicts the level to which two variables move in accordance with each another. If the two variables proceed to move in the same direction, then those variables have a positive correlation. If they move in contradictory directions, then they have negative correlation.

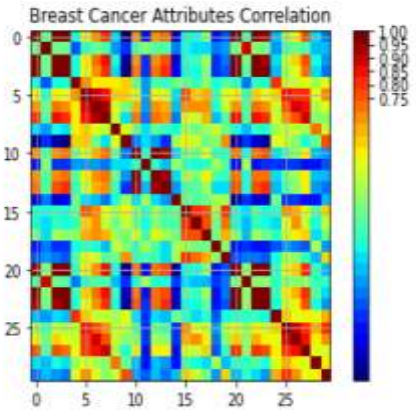


Figure2: Breast Cancer Attributes Correlation

In the resultant graph above, the red color around the diagonal indicates that features are corelated to one other. The yellow and green color patches depicts some common relation and blue color patch show negative relation. Finally, we split the data into predictor and target variables where predictor variable is a variable whose values will be used to predict the values of the target variable which is the outcome. It is comparably breaking into two sets. We will be using 20% of the test data. From the dataset, we will be able to evaluate and construct a version to derive if a given set of signs causes cancerSince we are not aware which one is the best. We do a spot review on some of the algorithms with default setting to get an early indication of how each them perform.

c. Cross validation

It is a method to analyze the auspicious models by segregating the exact sample into a set called training set to train the prototype and a set called test set to assess it. In cross validation, the primary sample is subdivided into equal size subsamples. From the subsamples a single subsample is maintained as the attestation data for examining the prototype and remaining subsamples are taken as training data. This validation technique is then performed again n times with each k subsamples. Their results are then combined to produce a single estimation.

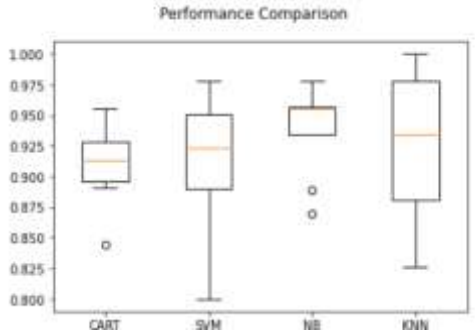


Figure 3: Performance comparison

Support Vector Machine has a unexpectedly bad performance in the situation

d. Estimation of algorithm on standardized input

The overall performance of few algorithm might be advanced if a systematized dataset is being used. The development is probably for all the models. I will be using pipelines that normalize the facts and construct the version for every fold within the cross-validation test harness. In that manner we will get a truthful conclusion of how these model with normalized data would possibly carry out on unseen data.

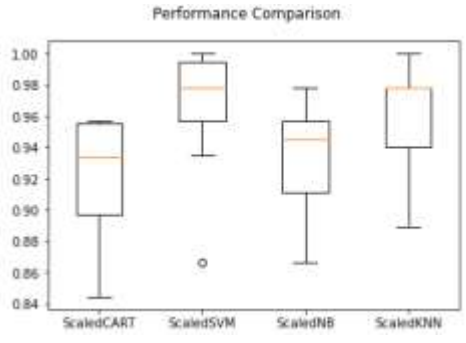


Figure 4: Performance comparison

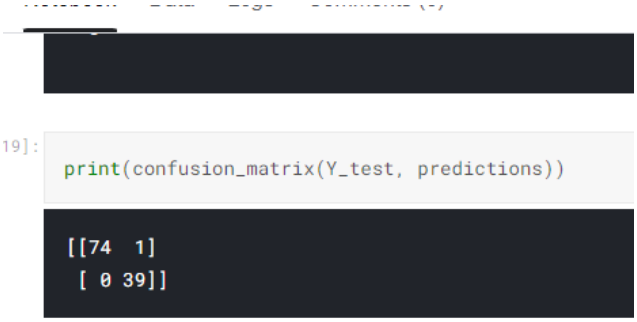
DOI: 10.5281/zenodo.6228080
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

There is a drastic change in the SVM after using the scaled data. Then we will tune the algorithm.

We implement a search model using the validation technique with a systematized replica of the training set. We apply a merger of values and the subsequent kernel types 'linear', 'poly', 'rbf' and 'sigmoid'. The most accurate layout was that of the SVM with RBF kernel, **C=1.5** and accuracy with **96.92%**.

e. Application of support vector classifier on dataset

Fit the SVM to the dataset



The output will be a confusion matrix which is a layout that enable visualization of the performance of an algorithm. It helps to calculate the performance of a issue. It have 2 dimensions, actual and predicted.

Accuracy is the estimation of exact prediction of the classifier, and it gives the basic data about the count of samples that are mixed-up.

Yet we attain an accuracy of 99% on the set. From the matrix, there is only 1 scenario that can be concluded as a mixed-up sample. The performance of this algorithm was expected to be high.

V. CONCLUSION

This study tries to resolve the hassle of automated recognition of breast cancer using a support vector machine. The current algorithm continues in exclusive levels. Many experiments were executed using the dataset. Experimental outcome concluded that the suggested system is more perfect than the existing one. It make sure that the proposed algorithm is important performance, productivity and quality and are vital in the medical world. With this conclusion we were capable to predicting a model with highest accuracy.

VI. REFERENCES

[1] Habib Dhahri, Eslam Al Maghayreh, Awais Mahmood, Wail Elkilani,and Mohammed Faisal Nagi, "Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms" Journal of Healthcare Engineering / 2019 / Article https://doi.org/10.1155/2019/4253641

[2] Mohammed Amine Naji, Sanaa El Filali, Meriem Bouhlal, EL Habib Benlahmar, Rachida Ait Abdelouhahid, Olivier Debauche,Breast Cancer Prediction and Diagnosis through a New Approach based on Majority Voting Ensemble Classifier,Procedia Computer Science,Volume 191,2021,Pages 481-486,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2021.07.061.

[3] "On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset" by Abien Fred M. Agarap, 7 February 2019

[4] "Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction", by Yixuan Li, Zixuan Chen October 18, 2018

[5] Analysis of Machine Learning Techniques for Breast Cancer Prediction" by the Priyanka Gupta and Prof. Shalini L of VIT university, vellore, 5 May 2018.

DOI: 10.5281/zenodo.6228080
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam