ISSN: 2584-0495 Vol. 2, Issue 12, pp. 1407-1413

International Journal of Microsystems and IoT

ISSN: (Online) Journal homepage: https://www.ijmit.org

Brest Cancer Predictor Using Machine Learning

Akhil Pratap Singh, Rohan Chauhan, Neha Aggarwal

Cite as: Singh, A. P., Chauhan, R., & Singh, R. K. (2024). Brest Cancer Predictor Using Machine Learning. International Journal of Microsystems and IoT, 2(12), 1407–1413. https://doi.org/10.5281/zenodo.15087000

© 2024 The Author(s). Published by Indian Society for VLSI Education, Ranchi, India

Published online: 24 December 2024

Submit your article to this journal:

Article views:

View related articles:

View Crossmark data:

DOI: https://doi.org/10.5281/zenodo.15087000

Full Terms & Conditions of access and use can be found at https://ijmit.org/mission.php

International Journal of Microsystems and IoT
Vol.2, Issue 12, pp.1407-1413 ; DOI: https://doi.org/10.5281/zenodo.15087000

Check for updates

Brest Cancer Predictor Using Machine Learning

Akhil Pratap Singh, Rohan Chauhan, Neha Aggarwal

Department of Computer Science Amity University, Uttar Pradesh

Corresponding Author e-mail: *rohanch23.work@gmail.com*

ABSTRACT

The problem of the Breast Cancer around the world among women becoming worsens each passing day. Curing of Breast Cancer disease is never easy for women. It is also become hard when women are in there 30's. After several research or experiments by our doctors and scientists, there is no 100% curable treatment for the cancer. In India, in every 4 minutes one women is diagnosed with breast cancer and in every 13 minutes one women died due to breast cancer. Researchers are predicting that by 2030 in India the most deaths among women will be due to breast cancer. We can also say that due to lack of accurate prediction models results in the difficulty for doctors to prepare for a treatment plan. So, we must develop a model which gives good accuracy with minimum errors in less time. Three types of algorithms SVM, AdaBoost, XgBoost which predict the breast cancer will be discussed. All these algorithms will be executed and conducted on JUPYTER Notebook platform. This work is done to predict the outcomes of different types of techniques and which technique has good accuracy.

Keywords:

Machine Learning, Breast Cancer , Xg Boost (Extreme Gradient Boosting) , Support Vector Machine (SVM), AdaBoost (Adaptive Boosting), KNN (K-Nearest Neighbor),CNN Convolutional Neural Network)

1. INTRODUCTION

A. OBJECTIVE OF THE PROJECT

The aim of the project is to solve the problem such as Which ML algorithm is appropriate for any feature of breast cancer & how to decrease the calculation computation. In this we use classification technique such as SVM (Support Vector Machines), AdaBoost (Adaptive Boosting), XgBoost (Extreme Gradient Boosting), CNN (Convolutional Neural Network), KNN (K-Nearest Neigbor), NaïveBayes, Random Forest.

B. BREAST CANCER

Breast Cancer is a type of cancer that starts in the breast. Breast Cancer mainly occurs in women's only, but sometimes men can get Breast Cancer. Early checking of Breast Cancer can help people to cure this disease.
Breast Cancer can start from different parts of the body. Breast Cancer starts when the cells in the breast change and their shape are formed of a sheet of mass cells i.e., called Tumor. Tumor is of two types i.e., Malignant & Benign.
Malignant, sometimes called Cancerous Tumor in the medical field means that cancer can grow quickly and can spread into different parts of the body and this process
is called metastasis [1-5].

© 2024 The Author(s). Published by Indian Society for VLSI Education, Ranchi, India

Benign Tumor is not as dangerous as Malignant, but doctors will check people very closely for changes in their body. Benign tumor means in which affected cells will grow but not spread in the body. There are 3 stages of cancer i.e., First, Second and Third. These stages describe how much breast cancer has grown or how much it spread. It spreads when the cancer cells mix with the blood or lymph system and then transfer to other parts of the body. The lymph is a part of our body's immune system. It is a network of lymph nodes ducts or vessels & organs that work together to collect and carry the lymph fluid through the body tissues to the blood. Also, most of the times breast cancer spread through lymph nodes and this phase is called metastatic or stage IV breast cancer and i.e., the most advanced stage of the cancer. But not all women who have cancer cells in their lymph nodes develop metastases, and there are some women's who have not cancer cells but in their lymph nodes metastases develop.

C. MOTIVATION FOR WORK

Breast Cancer is the most affected disease present in women worldwide. 246,660 of women's new cases of invasive breast cancer are expected to be diagnosed in the U.S during 2016 and 40,450 of women's death is estimated. The development in Breast Cancer and its prediction fascinated. The UCI Wisconsin Machine Learning Repository, Kaggle Breast Cancer Dataset attracted as large patients with multivariate

attributes were taken as sample set.

Data Mining use Machine Learning technology to boost the accuracy of previously built models due to the worrisome rise in breast cancer patients throughout the world. The major goal of the study is to handle irrelevant data and choose an acceptable model. The post-pre- processing activity involves evaluating the model using evaluation metrics and improving accuracy.

D. ABOUT MACHINE LEARNING

Machine Learning means implementing knowledge using an artificial system. Like human beings, we understand everything that we face or learn every day. Using ML, the computer can generate different ideas and can find any solution of the problem likewise we humans do. For example, if a child sees any object, then he/she can recognize that object in the same way we must design a model that predicts or recognize that this is that object. If that model is predicting or recognizing the same object, then it is a good ML model.

In ML, we linked the data with the computer, and it can do anything like making predictions, face recognition, recognition, handwriting and some other activities. We only must develop a good model for any prediction or recognition which gives you the correct accuracy (not 100% but if we make a model and it gives 96% or 97% accuracy then it can be good or perfect ML model) [6].

In present day, Machine Learning becomes the import part of our technology. It helps people work more efficiently and creatively. Now a days, ML is used in image processing, in stocks exchange, predicting prices or detection of error patterns etc. In ML, there are Features and Labels. We understand this by taking one example like cat and dog have ears, eyes, nose so these are the features. If we give these features to the ML model, then will model confuse that this is cat or dog? So here the model will fail. So, solving these types of problem we use, the self-learning algorithms like KNN, Random Forest, SVM, XgBoost, skLearn or many more. After predicting if the model says that it's a dog or cat i.e., we called label means model gives the tag to that object [7-10].

Machine learning is a crucial component of data science, a rapidly increasing field. To provide classifications or predictions and uncover crucial insights in data mining projects, algorithms are trained using statistical approaches. Ideally, the choices taken in response to these insights impact important growth metrics in applications and businesses. Data scientists will be more in demand as big data continues to develop and flourish. They will be expected to assist in determining the most important business queries and the data needed to address them.

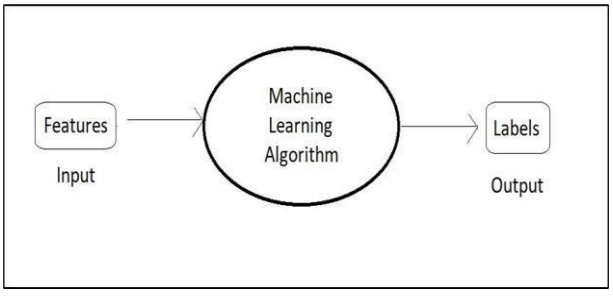


Fig 1. ML Diagram

Companies are constantly generating, exponentially, a large amount of data. Using this information, applying Machine Learning correctly is a great competitive advantage because high-value predictions can be obtained to make better decisions and carry out business actions. Machine Learning today is not how it was seen in the past. It was born because of pattern recognition and the theory that computers can learn without being programmed to perform specific tasks, at that time researchers interested in artificial intelligence wanted to see if computers could learn from data. The iterative aspect of machine learning is important because as models they are exposed to new data that can adapt independently. Machine learning models learn from the calculations above to produce repeatable results and decisions with a very high level of confidence. While many machine learning algorithms have been around for a long time, the ability to automatically apply complex mathematical calculations to large volumes of data is a recent development. The resurgence of interest in machine learning topics - machine learning - is due to the same factors that have made data mining and Bayesian analytics more popular today. Things like the volumes and varieties of data available, computational processing that is cheaper and more powerful, and growing accessible data storage[11-15].

Machine learning is basically divided into 2 types based on the way of learning:

1. Supervised learning: Supervised learning in total is about the learning process from data examples given in the previous label. Supervised learning requires labelled data to be able to perform data training, which is called the model. For example, by providing a lot of data in the form of photographs and suitable records, we can train the model to classify the individuals in the photos.

2. Unsupervised learning: Unsupervised learning is about modelling the input data without labels. With sufficient data, it is possible to find patterns and structures from data. The two most widely used tools in machine learning to learn from data alone are clustering and frugality.

E. SOFTWARE USED

For building the ML model we can use two languages Python and R. We will build our model in Python as python is easy to learn and easy to implement.

For downloading the Python, we only must follow some steps

First, we must download Python into our system. Write Google search (https://www.python.org/downloads/) bar Python install.

Click on python.org and download the latest version of Python. It will start downloading automatically. If you are using Mac OS or Linux, then go to the site and there will be option Python for Mac OS or Linux.

An open-source library called Python Pandas is described as offering high-performance data processing in Python. For both professionals and beginners, this tutorial is made.

After downloading, install Python and give the mission if

computer asked any.

So, we have Python in our computer. For implementation of ML, we must install some libraries like NumPy, Pandas, Matplotlib, sklearn, seaborn, xgboost.

Follow some steps to install these libraries:

a) Open PowerShell administrator (admin)

b) Write on this pip install sklearn matplotlib numpy pandas xgboost svm adaboost and it will start downloading.

c) After completing this write on PowerShell pip install jupyter notebook.

d) After downloading all these things write jupyter notebook on PowerShell and a new window will open.

e) Start implementing the code on the Jupyter Notebook.

2. LITERATURE REVIEW

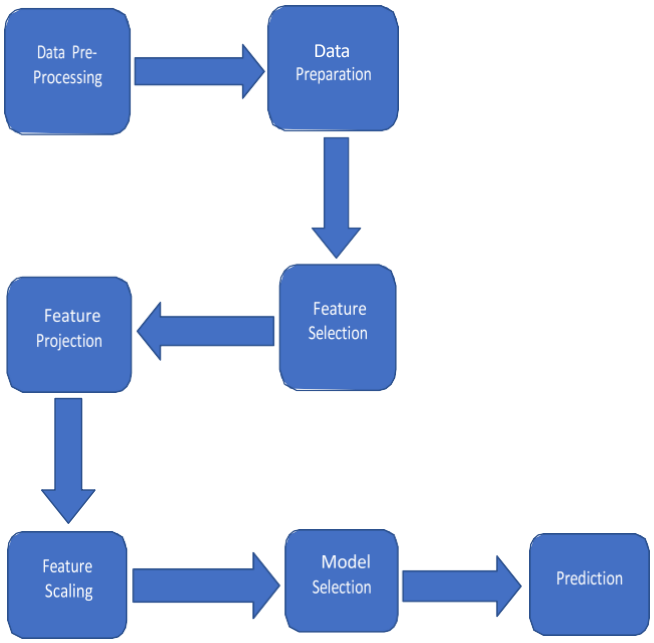
The main cause of Breast Cancer is changing and mutation in the DNA. It is also developed when the abnormal cells in breast divide & multiply. However, there are other factors through which it can increase or developing the breast cancer like Age (women which is older than 55 years can have more chances of this disease), Family or Genetics (if any of your family member like your parents, cousins or other close one who have been diagnosed with breast cancer, you have also been high chances to developing cancer), Smoking (around the world all the doctors always said that if you are using tobacco then there are always chances of having the cancer), Obesity (Increase of obesity can increase the chances of cancer), Hormone Replacement Therapy (people who use HRT has high chances of increase in breast cancer).
There are some treatments for breast cancer like Breast Cancer surgery, Chemotherapy, Radiation Therapy, Hormone therapy, etc [16-20].
We have used the algorithms SVM, AdaBoost and XgBoost.
Liu et al used the decision table that based on predictive models for survivability from breast cancer, resulting in that survivability of patients was 86.52%.
Tan and Gilbert explained the usefulness of employing the ensemble methods in classification of microarray data and they have some of the theoretical explanations on the performance of ensemble methods. After this, Tan and Gilbert give the suggestion to consider the ensemble machine learning methods for the process of classifying data of cancerous samples.
Chaurasiya and Pal differences the performance criteria of supervised learning classifiers like SVM-RBF kernel, RBF neural networks, Decision Tree (DT) and of simple classification and regression tree (CART). They are both differences to finding out the which classifier is best for breast cancer datasets. The experimental results are in favour of SVM-RBF kernel which shows that it is more accurate and better than other classifiers. The accuracy of SVM-RBF kernel is almost 97% when they provided the data of Wisconsin Breast Cancer data sets.

Table-1 Literature Review

No.	Year	Authors	Research Work	Algorithms Used	Accuracy
1.	2004	Delen at al.	Building a predictive model	AdaBoost	97.5 %
2.	2018	Chaurasia et. At	Understanding the raw data, locating the data to proper place	SVM	97.13%
3.	2019	Ch. Shravya	Building a SVM model	SVM	92.7 %
4.	2018	Sinthia et al.		Logistic Regression	94.2 %
5.	2017	Wang et al.	Suvivability of prediction using Logistix Regression	Logistic Regression	96.4 %
6.	2018	V Chaursiya and S Paul	Sorting of patient features by using the XgBoost algorithm.	Feature Selection in Static form	92.3 %
7.	2019	Akbugday		SVM	96.85%
8.	2018	Khourdifi et al.	Irrelevant attributes are deleted.	SVM	96.1 %
9.	2019	Mohana et al.	Helps in splitting the data.	Decision Trees	96.3 %
10.	2010	Al-hadidi	Optimizes the performance and cost function.	Logistic Regression	Greater than 93.7 %
11.	2016	Kibeom	Multiple Learners	AdaBoost	92.62%
12.	2013	Medjahed	Splitting the data	Logistic Regression	96.1 %

3. METHODOLOGY

A rough draft of the proposed work to follow in this study is as follows. It might change as the research progresses, but this flow chartgives an outline of the work.



Pre-Processing Data: The first thing we must do is to collect data for pre-processing and applying the Classification and Regression methods. Data pre-processing means it is a technique of data mining that involves the converting of raw data into an understandable format. Many researchers say real world data is not complete; it is fully inconsistent and contains so many errors. Because of this Data Pre-Processing gives solutions in these types of issues. For data pre- processing we have taken the data from Kaggle, or we can take data from UCI Machine repository for pre-processing. In our project, we collected the data of Breast Cancer samples from Kaggle which are Malignant and Benign. This data will be our training data set. Data Preparation: In this we must prepare the data to load into a suitable place and use it for machine learning training. We will put our data together and then we randomize the ordering [21].

Features Selection: In ML, it is also called as variable selection or attribute selection, which means the method of selecting a useful feature for building the ML model.

Feature Projection: C onverting the larger number of attributes to lower attributes. In this, both linear and non-linear reduction techniques can be used among the features in the data set.

Feature Scaling: In most of the data sets, there are so many features which have high magnitude, units and range. But now a days ML algorithms use Euclidian distance between two data points in their computations. In this stage our major aim is to bring all the features on the same magnitude.

Model Selection: Supervised Learning means that machines already know about any object or other feature. It involves the labelling of data sets so that algorithms get trained to easily classify any data or predict the future outcomes. Supervised Learning grouped in two categories, i.e., Classification and Regression techniques.

Classification Supervised Algorithm means predicting the labels after you trained the model after giving enough features and corresponding labels. It means the algorithm which predicts the specific thing int the data set & defines how this label should be for this data and this algorithm known as Classifier.

Regression means which predicts the specific value for any data set after recognizing the label or feature. So, the algorithm which predicts the continuous value using regression is called Regressor.

Types of Regression

a) Linear Regression

Linear regression fits a line (plane or hyperplane) to set of data; it is powerful, simple, and fast. (Microsoft, 2016). Linear regression belongs to the supervised learning category and is a regression since its output is continuous values. The goal of regression is to minimize the error between the approximate function and the value of the approximation. Indicate that when the result or the class is numeric and all attributes are numeric, linear regression is a natural technique to consider. Also, the idea is to express the class as a combination of the attributes, with determined weights:

$$x = w0 + w1a1 + w2a2 + \cdots + wkak$$

Where x is the class; $a1$, $a2$,…, ak are the attribute values; Y $w0$,
$w1$,…, wk are the weights. The weights are calculated from the data of the training. You need a way to express the values of training.

b) Logistic Regression

Logistic Regression is a linear regression technique used to classify linear and non-linear data. This model captures the vector of the variable and evaluates the coefficients or weights for each input variable and then predicts the class represented as the word vector. Because it is classified with a binary response, in machine learning the label must first be transformed with probability values 0 and 1, with 1 indicating success and 0 failings. However, the actual values that can be taken by 1 and 0 vary widely, depending on the research objectives. This model classifies data based on the probability of the variable occurring with a binary response with the equation:

$$p = \frac{exp\ (B0 + B1X)}{1 + exp(B0 + B1X)} = \frac{e^{B0+B1x}}{1+e^{B0+B1X}}$$

The probability transformation $\pi\ (x)$ is formulated by the following equation:

$$g(x) = ln\left(\frac{\pi\ (x)}{1 - \pi\ (x)}\right.$$

SVM

SVM is an amalgamation of existing classification theories such as Lagrange, kernel and margin hyperplane. SVM has a dividing function in the classification of the two classes

linearly but due to the increasing need for classification SVM was developed not only for linear classification. However, SVM can already do non-linear classification with a combination of kernels in the feature space (high dimensional space). SVM can also perform regression whose output is real numbers, or this algorithm is also called Support Vector Regression (SVR). By using SVR, classification prediction between several classes can be made. SVM is a set of supervised learning methods for classification, regression and outlier detection. The advantage of SVM is that it is effective in high-dimensional spaces and uses a subset of training points on the decision function (support vector). The drawback is that if the number of features is greater than the number of samples, the method gives less satisfactory performance and does not provide an estimated probability. The support vector machine forms a hyperplane or a set of hyperplanes in the infinite dimensional space used in classification and regression. The advantage of SVM is that it is effective in high dimensional spaces and different function kernels can be defined for decision functions. SVM is a machine learning system whose working principle uses Structural Risk Minimization (SRM). SRM aims to get the best dividing line (hyperplane) on the input space in the two-class classification. By measuring the margin on the dividing line and finding the vertex, it will be possible to find the best dividing line between the two classes. The distance in each of the two classes is called the margin and the pattern closest to the classification between the two classes is called the support vector. The learning stages in SVM are:

$$\sum_{} \alpha i - l \sum_{} \alpha i \alpha j y i y j \vec{x} . \vec{x} ..$$

Where i=12 $i.jl\, i \quad j$

Information:
yi= training data class (+ 1 / -1).
yj= training data class (+ 1 / -1).
xi= weight vector for the comment sentence.
xj= weight vector for the comment sentence

Unsupervised Learning is the other type of machine learning. In this, we have not provided the labels and features for the machine. In this machine is also not trained. That's why in unsupervised learning data becomes messy. Likewise, in supervised learning we have Classification and Regression here in Unsupervised Learning we have Clustering and Association. Clustering in ML means finding the inseparable groupings in the data. We put the similar data together in clustering. If we take an example of apples, banana and slices of banana full of basket, we expect that by using the clustering algorithm the machine would simply cluster all the fruits separately. It makes the cluster of apples on one side and bananas together on the other side. Association means to map the information which is based on the previous searches or activities. For example, if we place an order on Amazon or Flipkart for shoes, then due to this algorithm, Amazon or Flipkart will show more information or other types of shoes. So, this algorithm maps the current searches or information to the previous activities.

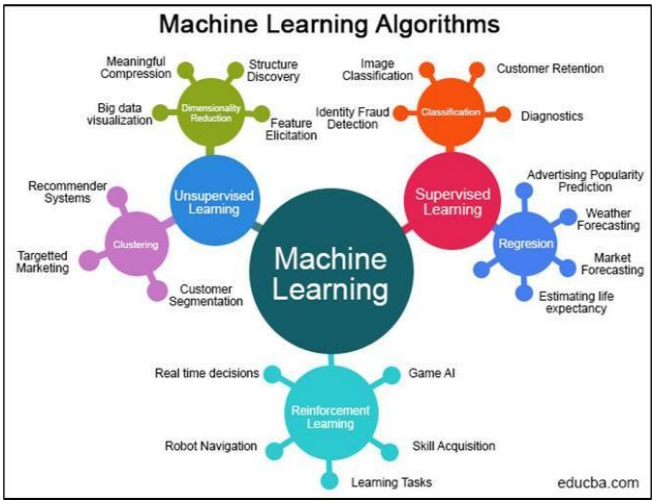


Fig 2. ML Algo

Prediction: ML uses the different types of algorithms to answer different types of questions. So, after building the model, the Prediction will be that stage where we get the answers to the questions.

4. RESULT AND EXPERIMENT ANALYSIS

We have implemented the ML model on JUPYTER notebook which contains different machine learning algorithms for pre-processing, classification, regression, clustering and association. The accuracies we are obtaining by using the Train and Test Method are in the table below:

Table 2- Accuracies (Train & Test Method)

Algorithms	***Accuracy (in %)***
SVM	96
XgBoost	98
KNN	96
CNN	97
Random Forest	96
AdaBoost	96
Decision Tree	95
Naïve Bayes	97

The accuracies we are obtaining by using the Cross-Validation method is in the table below:

Algorithms	***Accuracy (in %)***
AdaBoost	95.43
XgBoost	96.66
SVM	97.36

Random Forest	96.31
KNN	97.01
CNN	97.71
Naïve Bayes	93.67
Decision Tree	

Table 3- Accuracies (Cross-Validation Method)

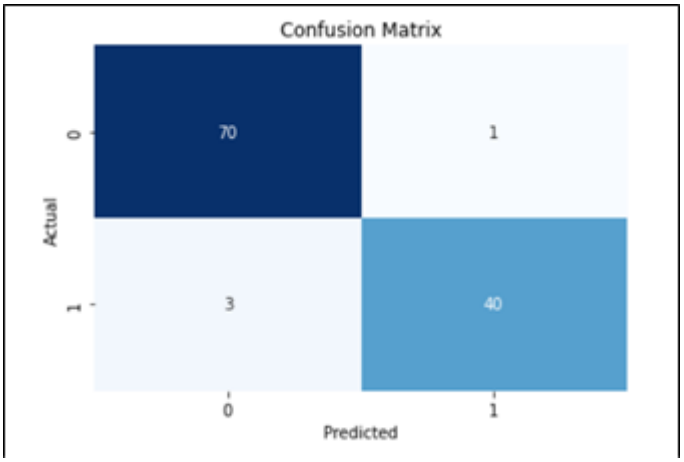


Fig 3. Random Forest

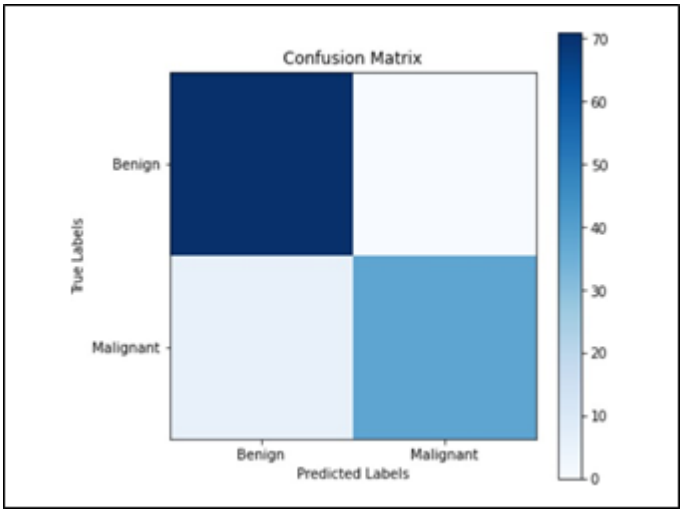


Fig 4. KNN

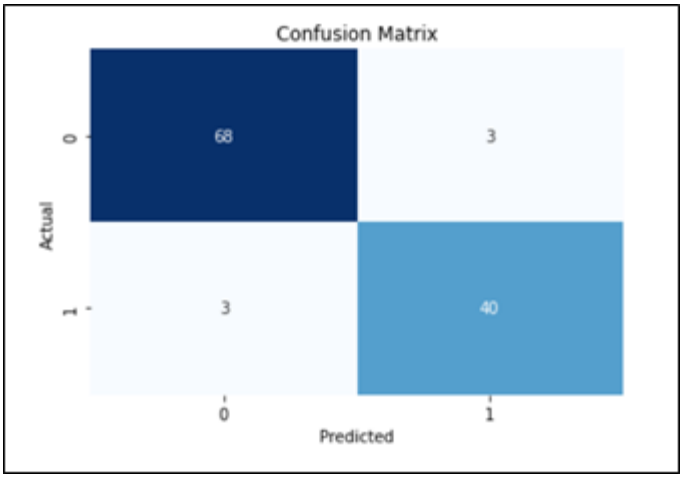


Fig 5. Decision Tree

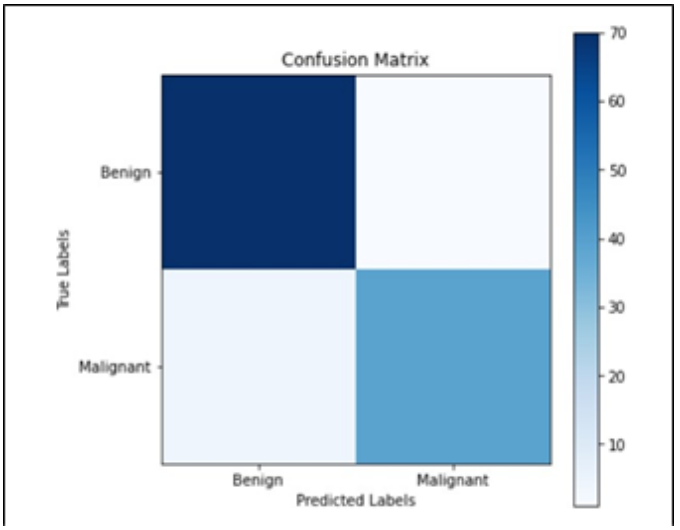


Fig 6. SVM

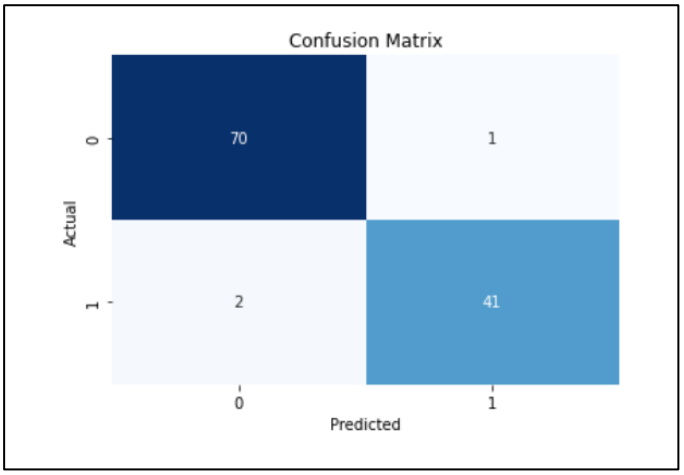


Fig 7. CNN

5 CONCLUSION AND FUTURE SCOPE

This model is based on Breast Cancer i.e., applied on the data sets which is taken from Kaggle and GitHub. We have taken the following steps to implement this model, i.e., Data Preprocessing, data preparation, Feature Selection, Feature Projection, Feature Scaling, Model Selection and Prediction. We have performed the data analysis by using the Python Pandas libraries. In the data preprocessing stage, we lowercased the document and removed the non-usable features such as id, unnamed. After the preprocessing we have selected the appropriate feature for building the model. To know how well the classifier is we have used the confusion matrix. The accuracy obtained by all the algorithms mentioned in the table. In conclusion we can also say that for small datasets we can use the K-Fold Cross- Validation Score Method and for large datasets we can prefer the Training and Testing method. Our model is not overfitted. Hence, we can say that our model is good.

The research which is done in future will be for the better performance of different classification techniques. We must improvise the classification techniques to give the high or better accuracy. In further research we will use the same model on the different- different data sets. We must decrease the error rate to improve the accuracy.

In future, we will also implement this model by using Deep Learning and put our time to improve the accuracies of the model by using all these algorithms which we have used.

The research which is done in future will be for the better performance of different classification techniques. We must improvise the classification techniques to give the high or better accuracy. In further research we will use the same model on the different data sets. We must decrease the error rate to improve the accuracy.

6. REFERENCES

1. Wang, D. Zhang and Y. H. Huang “Breast Cancer Prediction Using Machine Learning” (2018), Vol. 66, NO. 7.
2. Akbugday, B., et.al. (2019). Classification of Breast Cancer Data Using Machine Learning Algorithms, 2019 Medical Technologies Congress (TIPTEKNO), Izmir, Turkey,1-4.
3. Kaya, M., et.al. (2019). Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study. Tehnicki Vjesnik - Technical Gazette, 26(1), 149.
4. Chaurasia, V., et.al. (2014). Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability, IJCSMC, 3(1),10 – 22.
5. Delen, D., et.al. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. Artif.Intel. 34(1), 113–127.
6. Kavitha, R. K., et.al. (2014). Breast Cancer Survivability Using Adaptive Voting Ensemble Machine Learning AlgorithmAd boost and CART Algorithm, 3(1).
7. Sivasankari, R., et.al. Breast Cancer detection using PCPCET and ADEWNN, CIEEE’ 17(1),63-65.
8. Chaurasia, V., et.al. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis (FAMS 2016) 83(1),1064 – 1069.
9. Mishra, N., (2018). A Review of Breast Cancer Diagnosis in Mammography Images Using Deep Learning Techniques, 1(1).
10. Khourdifi, Y., et.al. (2018). Feature Selection with Fast Correlation-Based Filter for Breast Cancer Prediction and Classification Using Machine Learning Algorithms, 2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), Rabat, Morocco, 1-6.
11. Mohana, R. M., et.al. (2019).Lung Cancer Detection using Nearest Neighbor Classifier , International Journal of Recent Technology and Engineering (IJRTE), 8(2),S11.
12. Pravalika, K., et.al. (2019). Prediction of Breast Cancer Using Supervised Machine Learning Techniques, International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8(6).
13. Wang, H., et.al. (2015). Breast Cancer Prediction Using Data Mining Method, Proceedings of the 2015 Industrial and Systems Engineering Research Conference.
14. Bellaachia, A., et.al. (2020). Predicting Breast Cancer Survivability Using Data Mining Techniques © 2020 JETIR May 2020, Volume 7, Issue 5 www.jetir.org (ISSN-2349-5162) JETIR2005145 Journal of Emerging Technologies and Innovative Research (JETIR) www.jetir.org 24.
15. Kim, J., et.al. (2013). Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data, Journal of the American Medical Informatics Association,20(4), 613–618.
16. Khuriwal, N., et.al. (2018) . Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm," 2018 IEEMA Engineer Infinite Conference (eTechNxT), 1-5.
17. Amrane, M., et.al. (2018). Breast cancer classification using machine learning, 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), 1-4.
18. Desai, S. et. al. (2023). A Novel Technique for Detecting Crop Diseases with Efficient Feature Extraction. IETE Journal of Research, Taylor & Francis, https://doi.org/10.1080/03772063.2023.2220667
19. Patro KAK, Acharya B., **Nath V.**(2020). Secure, Lossless and Noise-resistive Image Encryption using chaos, Hyper-chaos and DNA sequence operation. IETE Technical Review, 37(3), 223-245. https://www.tandfonline.com/doi/full/10.1080/02564602.2019.1595751
20. Priyadarshi R., **Nath V.** (2019). A novel diamond-hexagon search algorithm for motion estimation. Microsystem Technologies, 25(12), 4587-4591, https://link.springer.com/article/10.1007/s00542-019-04376-5
21. Priyadarshi R, Soni S.K., Bhadu R, **Nath V.** (2018). Performance analysis of diamond search algorithm over full search algorithm” Microsystem Technologies, 24(6), 2529–2537, https://doi.org/10.1007/s00542-017-3625-0