

Comparison of Machine Learning Algorithms for Breast Cancer Prediction

Priyanshu, Dr. Vishal Verma, Neetu

MCA Student, Associate Professor, Asst. Prof.

Deptt. of Comp. Sc. and Application, Ch. Ranbir Singh University, Jind, Haryana, India

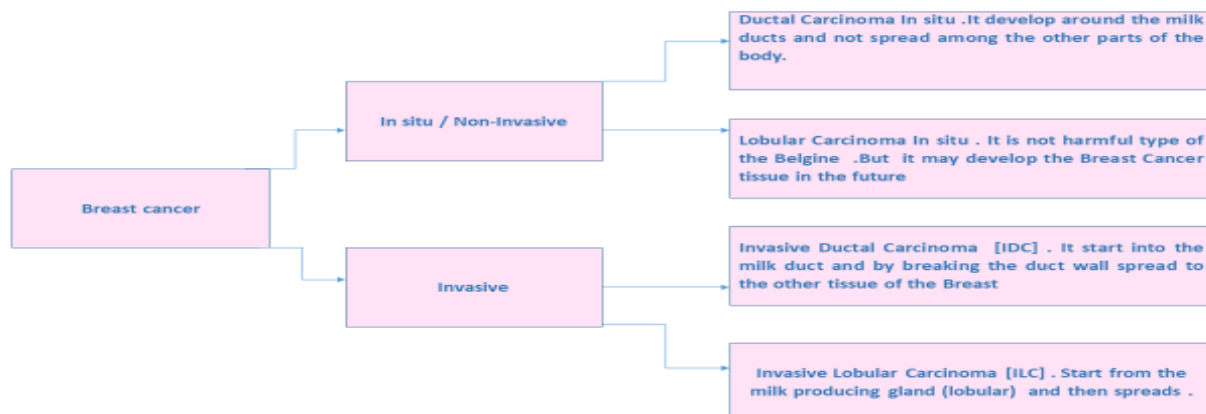
A eye catching point in healthcare sector breast cancer increment rapidly .Among cancer breast cancer is the most rapidly and worrisome in all other type of cancer like lung cancer, Bladder cancer, Skin cancer , Blood cancer ,Bone cancer , Pancreatic cancer , Immune system cancer[8].A discovery by the GLOBOCAN 2020 report find that among 10 million of death the 19.3 cases are cancer deaths cases in which female patient died are recorded 2.3 million which is caused by Breast cancer [1][4]. The diagnose of breast cancer is drawn-out-process with consumption of a lots of funds by physical examination . By using machine learning some model are used in this paper by accompaniment of WBC . Model used there are classification :- SVM, KNN . Ensemble techniques like RF , XGboost [4] . The result get after analysis is that XGBoost perform best among all of the algorithms where its accuracy is 0.9825 and recall is 0.9524

* ML = Machine learning , SVM = Support vector machine , K-NN= k-nearest neighbour , RF = Random Forest , WBC = Wisconsin Breast Cancer

1. Introduction

Breast cancer defined as unusual growth of cell in breast is either more or less as compare to cell in the other part of the body . Cancer are invase meaning a lumore may spread from breast to other part of body . Mainly there are two type of the cancer is Insitu and Invasive . Insitu are also called non-invasive . It do not spread to the other part of the body are invase meaning a lumore may spread from breast to other part of body .But Invasive may spread to the other parts of the body [14]. According to the report of WHO in 2022 most of the cancer are found female in 152 countries but the rate of Breast Cancer in men is 0.5 – 1.0 % nearly 80%[14.] According to the WHO report of which is released in 16 April 2026 declared that in-between the other type of cancer (lung cancer ,Bladder cancer , kin cancer , Blood cancer ,Bone cancer , Pancreatic cancer , Immune system cancer) breast cancer is the very noticeable disease analysis report. [12]. There are multiple whys for Breast cancer is Genetic Mutation , hormones , age , lifestyle , Radiations Exposure , Reproduction history [12]. The familiar

method for inquiry for making the difference in malignant and Benign breast cancer is Mammography (x- ray) ,ultrasonography (ultrasound) , Clinical breast cancer exam , Biopsies , Magnetic reasons images[1].



1.1 Motivation

GLOBOCAN 2020 report declare that among 10 million of death the 19.3 cases are cancer deaths cases in which among female patient death are recorded 2.3 million which is caused by Breast cancer[1][4]. If we detect the illness in early stage it help to reduce the rate of mortality in the females . But , in today era there are many techniques are developed which help to analysis the condition of patient and save a lots of time[8] . ML is that technique which help to make decision as soon as possible by using the supervised and unsupervised , reinforcement learning . Different model are applied in them . By using these model it aids to detect the disease in the early stage by which the life of the person is saved . ML is very helpful to detect the pattern and reasons of the illness causing . ML is very cost effective [1].

2. Literature review

The research done by Varsha Nemade et al. for breast cancer using WBC dataset from Kaggle . their they apply different algorithms on it which are RF, Adaboost , XGBoost , LR , SVM , K-NN, DT , NB .By these they discovered that among all other classification XGBoost and DT get highest accuracy of 97% , highest AUC of 0.999 by XGBoost.[1]In Breast Cancer Prediction by other Md. Milon Islam et al. doing researcher in algorithms [SVM , KNN , RF , LR , ARTIFICIAL NEURAL NETWORK(ANN)] by using the WBC dataset from Kaggle where they give decision on the bases of their accuracy , F1 score and precision also . Their they find highest accuracy among two algorithms which are SVM

accuracy is 97.14% , precision 95.14 % , F1 score is 0.977 . But, ANN have high accuracy ,precision , F1score is 98.57%, 97.82%,0.9890 .[2]The paper of Sara Laghmati et al.. CAD(computer aided diagnosis) system used by scholar to detect the malign and benign cancer . Their they use selection , PAC , Hypermeters , cross - validation to improve the accuracy compared to the work done on the past. they use classification ANN , KNN ,SVM, DT , RF , XGBoost , Adaboost and enabling method like LR , stacking , voting .use of accuracy , F1 score , precision , recall, include and mammography mass dataset and Wisconsin . their they find that XGBoost is the best for the Homographic dataset ,but for WBCD Adaboost and SLR mode have high recall of 95.35%.[3] Their they work on the deep ML . They work only on the Decision tree which give accuracy of the 83% only on by Mahendran Botlagunta et al..[4] The study on BREAST CANCER by investigator Somil Jain. Work on ensemble –based machine called ELRL-E using WBC dataset from Kaggle. they work on the Extra tree classifier, light GBM , Ridge classifier , linear Discriminate Analysis using a voting machine. They proposed that ELRL-E archive accuracy of 97.6% precision 96.4% , recall 100%, F1 score 98.1% .[5] In the paper wrote by AMREEN BATOOL , YUNG- CHEOL BYUN .There they use the WBC dataset where they applied different algorithms on it like SVM , NB , RF and many more but it is find that SVM have higher accuracy among all of them with 97.89% and error rate is 0 . 9 7 7 7 . [6]It is find that the researcher Youness Khourdifi et al. use the WBC dataset and use algorithms like SVM,RF,K-NN , Naive Bayes where as the SVM perform better in all of them with the high accuracy of 97.9% which help in better prediction of breast cancer for the patient and the doctor[7].In the paper written by VIBITH A S mainly use the two algorithm like **GBDTMO (Gradient Boosting Decision Tree with Mayfly Optimization)** and **DLMN-EHO (Deep Learning Model with Elephant Herding Optimization)** . Where the accuracy of both model is same which is 96.35% but the recall of the GBDTMO is 98.52% is better where as the recall rate of DLMN-EHO is 97.32%.[8]In the paper written by the Tharunya Arravalli et al. .It is decide that among all the other model RF give the best F1 –score which is 84% where the F1-score given by the stacked ensemble model is 83% which is lower then the RF .[9]It is analysis the paper written by Sharifah Nurulhikmah Syed Yasin et al. that among all the model like Adaboost and XGBoost .RF perform good with accuracy 91.23% , precision 90.70% , recall 86.67% , F1-score 88.89% by using the Wisconsin (Diagnostic) Data Set.[10]

3. Methodology

3.1 Dataset Description

The first step to start any research to collect the data for using in the research. In this paper secondary data which is collected from the kaggle . The dataset we take from kaggle are Wisconsin Breast Cancer [WBC] which are abstract from university of Wisconsin[13] There name of the columns are:-mean radius, radius error, worst radius; mean texture, texture error, worst texture; mean perimeter, perimeter error, worst perimeter; mean area, area error, worst area; mean smoothness, smoothness error, worst smoothness; mean compactness, compactness error, worst compactness; mean concavity, concavity error, worst concavity; mean concave points, concave points error, worst concave points; mean symmetry, symmetry error, worst symmetry; and mean fractal dimension, fractal dimension error, worst fractal dimension ,unnamed. There are 32 columns and 10 rows.

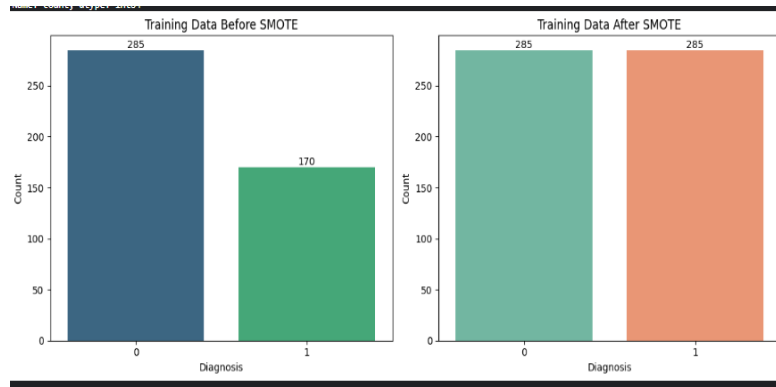
3.3Data Processing

In this step missing value is handle , remove the unnecessary or unwanted columns , checking the class imbalance and correct it . If there are any categorical data then change it into the binary form because machine can't understand the human read able language . Extract the meaningful feature by analysis the features dependency that actually help to analysis the result.

- First we check the number of columns and rows .Find the missing values . In this dataset there are one column name unnamed have no value .
- Then remove the columns which are not necessary for the calculation like id , unnamed .Change the categorical data into numerical data .

3.4 Training and Testing Phase

After applying the two steps the training and testing come . In this step the cleaned dataset is split into the two categories . The 80% of the dataset is use for training the data we collected and 20% is use to test to look that the result give by how much predicted result are true.



Class imbalanced

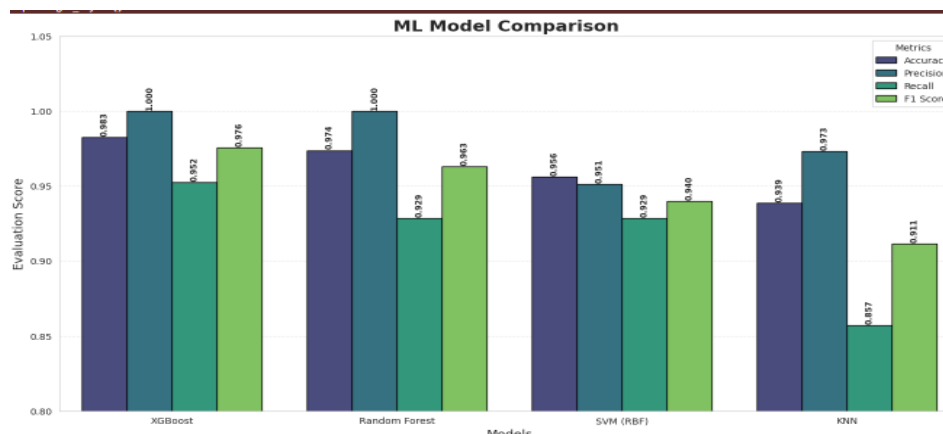
There it is analysis that class are balanced or not . If the class is not balanced then we balanced it by using the smooth method . In WECD there are imbalanced class Malignant are 170 sample where as benign have 285 sample .So, there use the smooth method to balanced it .

3.5 Model selection and Applied it .

To extract the correct result different models are applied . In this paper classification models (SVM ,KNN) are used and Ensemble model (XGBoost , RF) applied in it . Classification model divide into the form of class and ensemble model combine them to get the better result .

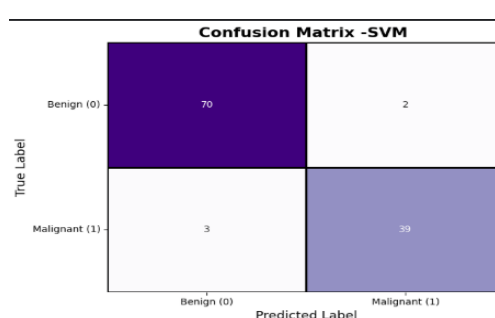
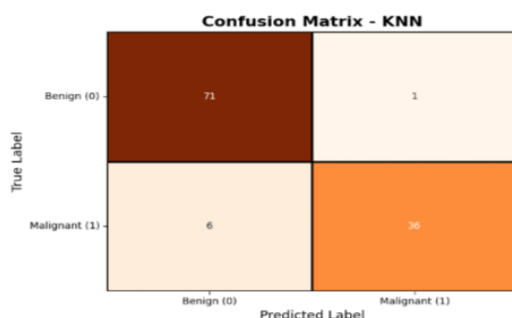
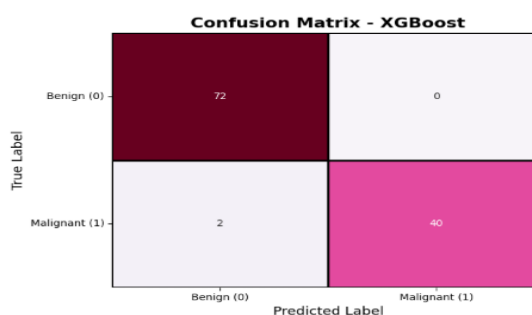
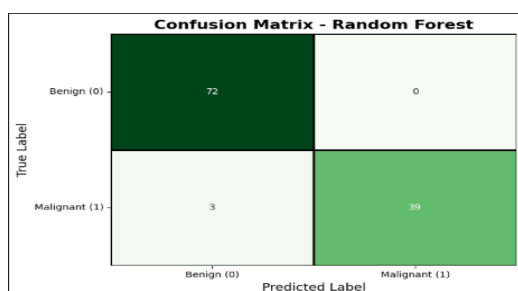
- **SVM**- It is used for both classification and regression . In SVM classified data for nearest hyperplane . There are two type in it which are linear and non-linear
- **RF** – RF stand for Random Forest . It is a collect of Decision Tree . RF randomly split the data and make tree for them and combine them to get the result by voting(classification) or averring (regression) .
- **XGBoost** – It stand for Extreme Gradient Boosting . It make the tree but it also correct the error done previously
- **KNN** – It working to predict the results to its nearest similar point as compared to the other point..

3.6 Comparison of Graph and confusion matrix



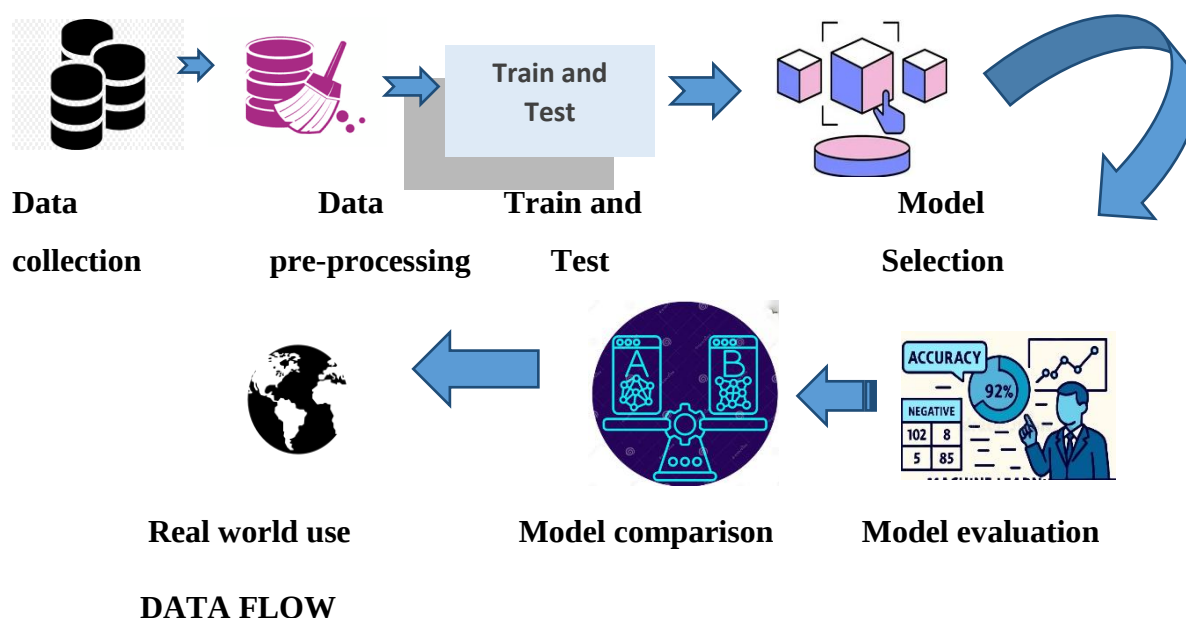
Comparison Graph

In this graph above tell that the best accuracy , f1 score, recall among all the model is obtained by the XGBoost which is 98% , 0.9756,0.9524. whereas the RF perform slightly less than the XGBoost by giving accuracy(97%) and recall(0.9286) . Meanwhile, worst performance is done by the KNN is 93% accuracy and its recall rate is 0.8571 . But, the SVM production of accuracy is better than KNN but less than RF. In SVM the use of RBF kernel is used and accuracy(95%) , precision (0.9512), recall(0.9286) , f1 score (0.9384)in not better than the RF , XGBoost



The conclusion achieved after analysis the confusion matrix that according to recall XGBoost only miss the 2 case of malignant meanwhile the SVM and RF miss 3 case , KNN

have 6 missing case . According to accuracy identification that out of 114 case XGBoost only made 2 incorrect prediction this type of similar work is done by RF which only predicted wrong 3 case . The miserable performance is done by the KNN which wrongly predicted the 107 case out of 114 case . SVM miss the 109 case out of 114 which means that it miss the 5 case. When prediction mean how many predicted case of malignant are true(false positive) in this area KNN tell that a patient have malignant but it tell it benign . RF & XGBoost not miss any case their . while the SVM have 2 wrong telling case which have higher number of wrong case in all the other models.F1 score is used to balance between the precision and recall. XGBoost have the best F1-score . As compare to the other model KNN have less f1-score.



4. Conclusion and Future scope

4.1 Conclusion

Breast Cancer have the major life threatening disease among the female . It is found for the among the Malaysian women it is the most rapidly increase disease [10]. By analysis the dataset and after apply models it is find that XGBoost give the highest accuracy(98%) and high recall rate(95%).

4.2 Future work

- Implement of big data .
- To identify a model which handle both outlier and robust together .

- Use image with data.
- The use of deep learning

Reference

- [1]Varsha Nemade and Vishal Fegade. “Machine learning techniques for breast cancer prediction”. *Procedia Computer Science*, 218:1314–1320, 2023.
- [2]Md. Milon Islam, Md. Rezwanul Haque, Hasib Iqbal, Md. Munirul Hasan, Mahmudul Hasan, and Muhammad Nomani Kabir.” Breast cancer prediction: A comparative study using machine learning techniques”. *SN Computer Science*, 1:290, 2020.
- [3]Sara Laghmati1 Soufiane Hamida , Khadija Hicham ,Bouchaib Cherradi , Amal Tmiri. “An improved breast cancer disease prediction system using ML and PCA” Received: 31 March 2023 / Revised: 10 July 2023 / Accepted: 4 September 2023 Published online: 25 September 2023
- [4]Mahendran Botlagunta, Madhavi Devi Botlagunta, Madhu Bala Myneni, D. Lakshmi, Anand Nayyar, Jaithra SaiGullapalli, and Mohd Asif Shah.” Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms”. *Scientific Reports*, 13, 2023.
- [5]Somil ,Jain and Puneet Kumar. Prediction of breast cancer using machine learning. *Recent Advances in Computer Science and Communications*,13(5):901–908, 2020.
- [6]Amreen Batool and Yung-Cheol Byun. Toward improving breast cancer classification using an adaptive voting ensemble learning algorithm. *IEEE Access*, 12, 2024.
- [7].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)*, volume 1, 2018.
- [8] Vibith A. S. Machine Learning Models for Breast Cancer Prediction in the Early Stages. Ph.d. dissertation, Anna University, Chennai, India, February 2025.
- [9] pathila, D. Cenitta, Rajagopala Chadaga, and K. S. Swathi. Detection of breast cancer using machine learning and explainable artificial intelligence. *Scientific Reports*, 15:26931, 2025.
- [10] Sharifah Nurulhikmah Syed Yasin, Aiman Azhar, and Rajeswari Raju. Breast cancer prediction: A random forest-based system with expert validation. *Pertanika Journal of Science and Technology*, 33(S3):157–180, 2025
- [11] Anas Bilal, Ali Alkhatlan, Faris A. Kateb, Alishba Tahir, Muhammad Shafiq, and Haixia Long. A quantum-optimized approach for breast cancer detection using squeeze-net-svm. *Scientific Reports*, 15:3254, 2025

[12]American Cancer Society. Breast cancer early detection and diagnosis. CA:A Cancer Journal for Clinicians, 70:7–30, 2020.C. Tan and J. Lee. Physical Examination Techniques in Oncology. Springer,New York, NY, USA, 2019. World Health Organization. Breast cancer: Early detection, 2021. Accessed: 2026-04-03.

[13][https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data`](https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data)

[14]<https://my.clevelandclinic.org/health/diseases/3986-breast-cancer>