Breast Cancer Detection Using Machine Learning

Santhosh M, Risha kesan S.R, Shanmughavel A.M,
Dr.S.Harihara Gopalan, Assoc.P/CSE

1. ABSTRACT

Women are seriously threatened by breast cancer with high morbidity and mortality. The lack of robust prognosis models results in difficulty for doctors to prepare a treatment plan that may prolong patient survival time. Hence, the requirement of time is to develop the technique which gives minimum error to increase accuracy. Four algorithm SVM, Logistic Regression, Random Forest and KNN which predict the breast cancer outcome have been compared in the paper using different datasets. All experiments are executed within a simulation environment and conducted in JUPYTER platform. Aim of research categorises in three domains. First domain is prediction of cancer before diagnosis, second domain is prediction of diagnosis and treatment and third domain focuses on outcome during treatment. The proposed work can be used to predict the outcome of different technique and suitable technique can be used depending upon requirement. This research is carried out to predict the accuracy. The future research can be carried out to predict the other different parameters and breast cancer research can be categorises on basis of other parameters.

Keywords — Breast Cancer, machine learning, feature selection, classification, prediction, KNN , Random Forest, ROC.

2. INTRODUCTION

The second major cause of women's death is breast cancer (after lung cancer). 246,660 of women's new cases of invasive breast cancer are expected to be diagnosed in the US during 2016 and 40,450 of women’s death is estimated. Breast cancer is a type of cancer that starts in the breast. Cancer starts when cells begin to grow out of control. Breast cancer cells usually form a tumour that can often be seen on an x-ray or felt as a lump. Breast cancer can spread when the cancer cells get into the blood or lymph system and are carried to other parts of the body. The cause of Breast Cancer includes changes and mutations in DNA. There are many different types of breast cancer and common ones include ductal carcinoma in situ (DCIS) and invasive carcinoma. Others, like phyllodes tumours and angiosarcoma are less common. There are many algorithms for classification of breast cancer outcomes. The side effects of Breast Cancer are – Fatigue, Headaches, Pain and numbness (peripheral neuropathy), Bone loss and osteoporosis. There are many algorithms for classification and prediction of breast cancer outcomes. The present paper gives a comparison between the performance of four classifiers: SVM , Logistic Regression , Random Forest and kNN which are among the most influential data mining algorithms. It can be medically detected early during a screening examination through mammography or by portable cancer diagnostic tool. Cancerous breast tissues change with the progression of the disease, which can be directly linked to cancer staging. The stage of breast cancer (I–IV) describes how far a patient’s cancer has proliferated. Statistical indicators such as tumour size, lymph node metastasis, and distant metastasis and so on are used to determine stages. To prevent cancer from spreading, patients have to undergo breast cancer surgery, chemotherapy, radiotherapy and endocrine. The goal of the research is to identify and classify Malignant and Benign patients and intending how to parametrize our classification techniques hence to achieve high accuracy. We are looking into many datasets and how further Machine Learning algorithms can be used to characterize Breast Cancer. We want to reduce the error rates with maximum accuracy. 10-fold cross validation test which is a Machine Learning Technique is used in JUPYTER to evaluate the data and analyse data in terms of effectiveness and efficiency.

3. MOTIVATION

As of 2019, on average, 1 in 8 U.S women (approx. 12%) would develop invasive breast cancer at some point during her life. 5-year survival rate for breast cancer is 100% with early detection and 15% with late detection (UK Cancer research) . Machine learning (ML) techniques play a key role in healthcare in recent years. In the case of breast cancer, machine learning techniques can be used to distinguish between malignant and benign tumours for enabling early detection. Most ML based applications focus on large data sets citing ML's ability to handle big data. However, from a user's perspective most users have access to publicly available small data sets. Thus, it is interesting to analyse if the traditional non complex basic ML algorithms can achieve high accuracy classifications using small datasets.

4. RELATEDWORK

A. Turgut Machine learning procedure compared with SVM, KNN, DT, Logistic Regression, Random Forest, ADA Boost. In this various method checked and conclude that highest efficiency is 89% of random forest.

B. Narasingarao.M presents a survey of the work conducted to detect breast cancer using with different algorithm and conclude the efficiency of algorithm.

C. Junaid Ahmed achieved 84.21% accuracy by using Adaptive Reasoning Theory, the Wisconsin data set was used, that contains 569 rows of data, and also contains 32 attributes.

D. Nithya [13] applied the three categorizing methods such as Decision Tree, k-Nearest Neighbour, and Naïve Bayes for the different datasets. The authors also inspect the evaluation metrics of error rate. The implementation was focused on a type of attribute of a dataset.

E. Shilpa M and C. Nandini [19] implemented the algorithm using python and tested the same using dataset and achieved an accuracy of 94.74 and also reduces the time taken.

F. Hafizah [2] compared SVM and ANN using four different datasets of breast cancer. The researchers have demonstrated that SVM was better than ANN in performance and result.

G. S. Gc [1] worked on extracting features including variance, range, and compactness. They used SVM classification to analyse the performance. Their findings showed the highest variance of 95% and compactness 86%. According to their results, SVM can be considered as an appropriate method for Breast Cancer Prediction.

5. PROPOSED METHODOLOGY

Phase 1 - Pre-Processing Data

The first phase we do is to collect the data that we are interested in collecting for pre-processing and to apply classification and Regression methods. Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real world data is often incomplete, inconsistent, and lacking certain to contain many errors. Data pre-processing is a proven method of resolving such issues. Data pre-processing prepares raw data for further processing. For pre-processing we have used standardization method to pre-process the UCI dataset. This step is very important because the quality and quantity of data that you gather will directly determine how good your predictive model can be. In this case we collect the Breast Cancer samples which are Benign and Malignant. This will be our training data.

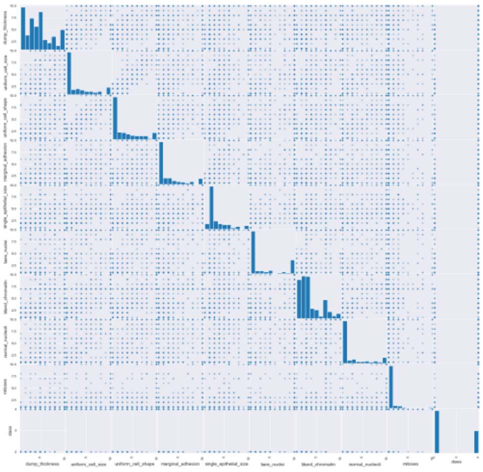
Phase 2 - Data Preparation

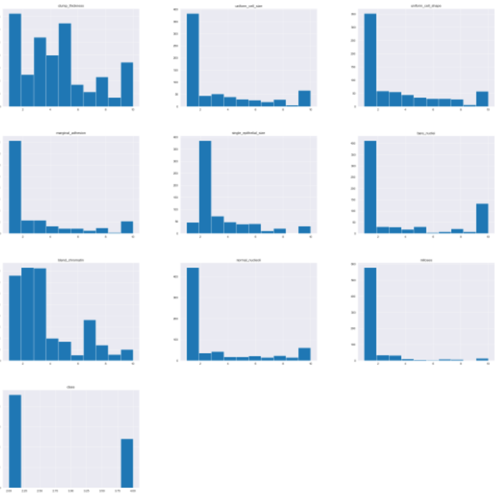
Data Preparation, where we load our data into a suitable place and prepare it for use in our machine learning training. We'll first put all our data together, and then randomize the ordering.

Phase 3 - Data Visualization

We are going yo Visualize our numeric data with respect to Two categories 1)Benign 2) Malignant

Plot histograms for each variable Create scatter plot matrix





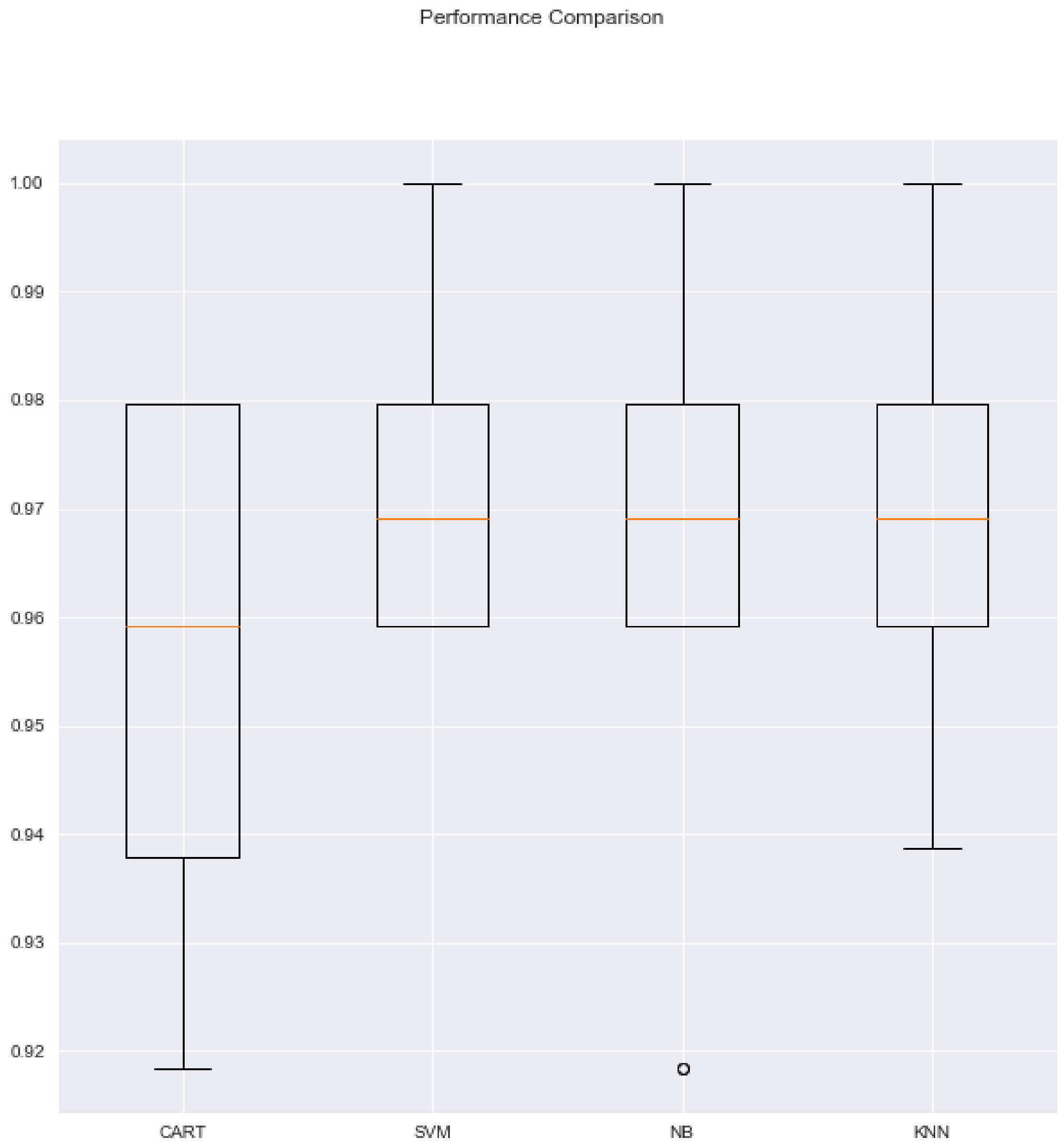
Phase 5 - Feature Scaling

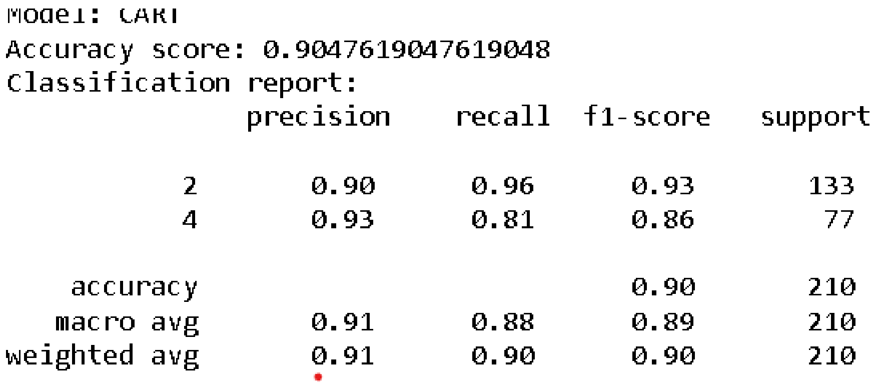
Most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Euclidian distance between two data points in their computations. We need to bring all features to the same level of magnitudes. This can be achieved by scaling.

Phase 6 - Model Selection

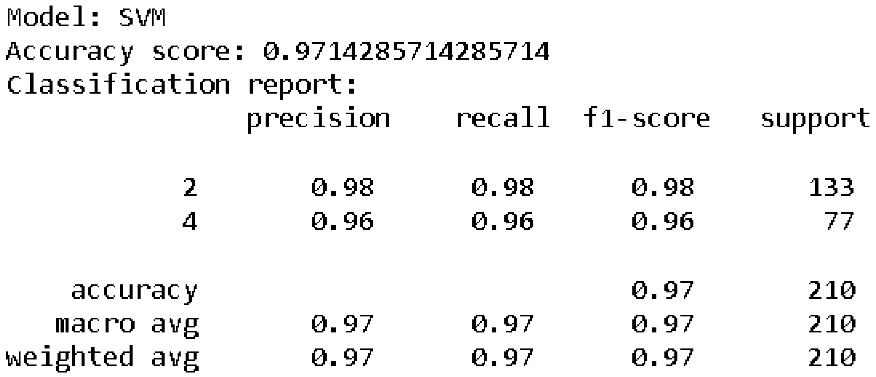
We used Jupyter lab and Anaconda prompt as a Coding platform and get a prediction output from the Flask in Local Server. Our Methods Includes Supervised Learning Algorithms and Classification Techniques like Support Vector Classifier (SVM), Random Forest, Naïve Bayes, Decision Tree, KNN, Adaboost and XGboost. Dataset contains features which highly vary in units and magnitudes. So, it is required to bring all features to the same level of magnitudes. We did that by using Standard Scaling in SKLearn.

Model selection is the most important step in Machine Learning. Machine Learning algorithms can be classified as: supervised learning and unsupervised learning. For Our project, we only need supervised learning. We used all Methodologies to Predict the result and Noted their Accuracy.

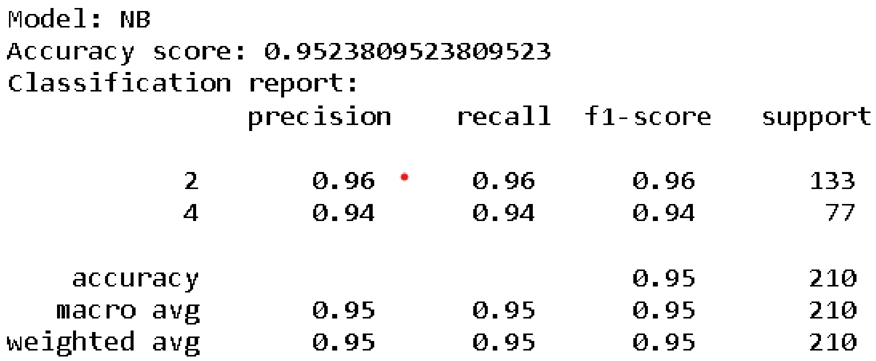




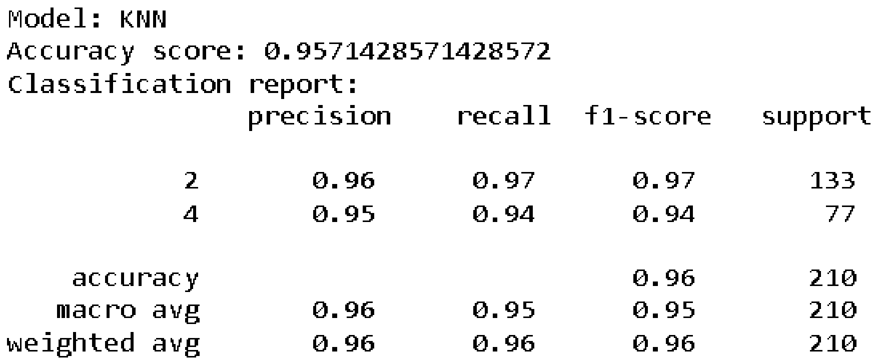
```
Model: CART
Accuracy score: 0.9047619047619048
Classification report:
               precision    recall  f1-score   support

           2       0.90      0.96      0.93       133
           4       0.93      0.81      0.86        77

    accuracy                           0.90       210
   macro avg       0.91      0.88      0.89       210
weighted avg       0.91      0.90      0.90       210
```



```
Model: SVM
Accuracy score: 0.9714285714285714
Classification report:
               precision    recall  f1-score   support

           2       0.98      0.98      0.98       133
           4       0.96      0.96      0.96        77

    accuracy                           0.97       210
   macro avg       0.97      0.97      0.97       210
weighted avg       0.97      0.97      0.97       210
```



```
Model: NB
Accuracy score: 0.9523809523809523
Classification report:
               precision    recall  f1-score   support

           2       0.96      0.96      0.96       133
           4       0.94      0.94      0.94        77

    accuracy                           0.95       210
   macro avg       0.95      0.95      0.95       210
weighted avg       0.95      0.95      0.95       210
```



```
Model: KNN
Accuracy score: 0.9571428571428572
Classification report:
               precision    recall  f1-score   support

           2       0.96      0.97      0.97       133
           4       0.95      0.94      0.94        77

    accuracy                           0.96       210
   macro avg       0.96      0.95      0.95       210
weighted avg       0.96      0.96      0.96       210
```

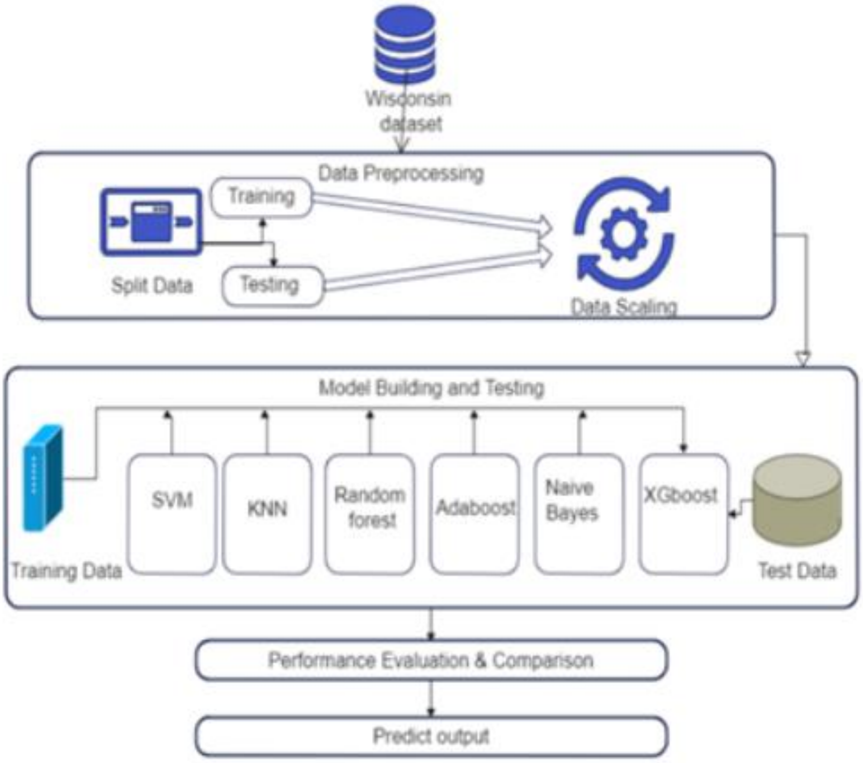
From the output abstained we select SVM as a model

Phase 7 -Confusion Matrix

Confusion Matrix is used for evaluating the performance of a classification model. The Matrix compares the actual target values with predicted values by the machine learning model. It shows the ways in which your classification model gets confused when it makes predictions

6.PROPOSED SYSTEM ARCHITECTURE

As Shown in diagram, we first Uploaded dataset From Wisconsin Breast Cancer Dataset. After that We did Preprocessing to the data and applied Machine Learning Models, which is used in this project to predict Breast cancer.



7.CONCLUSION AND FUTURE WORK

We can notice that SVM takes about 0.07 s to build its model unlike k-NN that takes just 0.01 s. It can be explained by the fact that k-NN is a lazy learner and does not do much during training process unlike others classifiers that build the models. In other hand, the accuracy obtained by SVM (97.13%) is better than the accuracy obtained by C4.5, Naïve Bayes and k-NN that have an accuracy that varies between 95.12 % and 95.28 %. It can also be easily seen that SVM has the highest value of correctly classified instances and the lower value of incorrectly classified instances than the other classifiers. After creating the predicted model, we can now analyse results obtained in evaluating efficiency of our algorithms. SVM and C4.5 got the highest value (97 %) of TP for benign class but k-NN correctly predicts 97% of instance that belong to malignant class. The FP rate is lower when using SVM classifiers (0.03 for benign class and 0.02 for malignant class), and then other algorithms follow: k-NN, C4.5 and NB. From these results, we can understand why SVM has outperformed other classifiers In summary, SVM was able to show its power in terms of effectiveness and efficiency based on accuracy and recall.

8.REFERENCE

6. S. Gc, R. Kasaudhan, T. K. Heo, and H.D. Choi, “Variability Measurement for Breast Cancer Classification Mammographic adaptive and convergent systems (RACS), Prague, Czech Republic, 2015, pp. 177–182.
7. S. Hafizah, S. Ahmad, R. Sallehuddin, and N. Azizah, “Cancer Detection Using Artificial Neural Network and Support Vector Machine: A Comparative Study,” J. Teknol, vol. 65, pp. 73–81, 2013.
8. A. T. Azar, and S. A. El-Said, “Performance analysis of support vector Neural Compute. Appl., vol. 24, no. 5, pp. 1163–1177, 2014.
9. machines classifiers in breast cancer mammography recognition,” Neural Comput. Appl., vol. 24, no. 5, pp. 1163–1177, 2014.
10. C. Deng, and M. Perkowski, “A Novel Weighted Hierarchical Adaptive Voting Ensemble Machine Learning Method for Breast Cancer 2015.
11. Z. Jiang, and W. Xu, “Classification of benign and malignant breast cancer based on DWI texture features,” ICBCI 2017 Proceedings of the Iinternational Conference on Bioinformatics and Computational Intelligence 2017.
12. R. Jegadeeshwaran and V. Sugumaran (2013) Comparative study of decision tree classifier and best first tree classifier for fault diagnosis of automobile hydraulic brake system using statistical features, Measurement, vol.46, pp.3247–3260.
13. Ajith Abraham (2005), Artificial neural networks, Nature & scope of AI techniques, vol.2, pp.901-908.
14. Jennifer Listgarten, Sambasivarao Damaraju, Brett Poulin, Lillian Cook, Jennifer DuFour, Adrian Driga, John Mackey, David Wishart, Russ Greiner and BrentZanke (2004), Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms, Clinical Cancer Research, vol.10, pp.2725- 2737.
15. Jaree Thongkam, Guandong Xu and Yanchun Sang (2008), Breast cancer survivability via AdaBoost algorithms, Health data and knowledge management, vol.80.