IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

BREAST CANCER PREDICTION

[1]Prof.Shivakumar.M, [2]Kokila R, [3]Likitha B S, [4]Tharun N, [5]Adishesha.R

[1] Assistant Professor, [2] Student, [3] Student, [4] Student, [5] Student

Department of Information Science and Engineering,
Cambridge Institute of Technology, Bangalore, India

Abstract: This final-year project aims to analyse and detect Breast Cancer . Womens are highly suffered from breast cancer,with huge medical problems caused by a treatment(Morbidity) and destined to die(Mortality). Due to lack of prediction models the accuracy of prediction or detection of breast cancer if difficulty. Because of this the prediction time and the patient duration for sustaing time need to be prolonged. Hence to predict early the technique or model is designed to give prediction accuracy exactly. In this SVM, DT, GaussianNB and KNN are the four algorithms used to predict breast cancer results compared with large and different datasets. The proposed model is selected to predict the result of many techniques and correct technique is used depending upon the treatment. This model is based on getting and future studies can be done to predict other methods it can be categorised on basisi of other methods.

Keywords –Breast Cancer, Machine learning, Support Vector Machine, Decision Tree, KNN and GaussianNB.

I. INTRODUCTION

Breast Cancer is the second leading cause of the cancer death in the women, second only to the lung cancer. Reason for breast cancer is simply being a women, though breast cancer is seen in men, the disease 100 times more common in women. Even though men can also get breast cancer. In the year 2017, the American cancer society estimates 2,470 new cases of massive breast cancer is predicted in men in the U.S. A women has one in eight chance of being with breast cancer. Most of the women who are getting breast cancer is not having any heredity problems. Breast cancer is one of the most prevalent life- threatening diseases which is affecting individuals, women globally. In this early detection and personalized intervention is very important. Breast cancer is one of the type of cancer that is seen in breast. Cancer starts when abnormal or uncontrolled growth of cells are seen. Breast cancer cells are those which grown out of control form a tissues and results in a tumour that can be seen on x-ray or felt as a lump or un-sized breast size. The side effects of breast cancer are – fatigue, headache, pain, numbness, bone loss, osteoporosis, other blood diseases etc. There are many different algorithms for prediction and classification of breast cancer results. This paper is mainly focused on the performance of four algorithms: Support Vector Machine(SVM), Decision Tree(DT), GaussianNB, KNN which are most Regression and classifier algorithms. It is early detected by doctors through a screening examination of mammography or portable cancer diagnostic tool. To prevent spreading of cancerous cells or tissues to other parts is stopped by undergoing surgery, chemotherapy, radiotherapy and endocrine. The final result of this research is to identify or classify the two different tumours and intention is how to parameterize our classification methods to achieve high accuracy. On classifying malignant and benign tumour or patients we can come to know that the patient with malignant tumour is having a breast cancer and the patient with bengin is not having a cancer or dangerous cells. Early and accurate detection or diagnosis of breast cancer could improve the clinical and patient survival results. Thus, this methods are important for detecting early signs of breast cancer. Mammography is the picture scanning technique for early detection of breast cancer, in this accuracy was not high. So, Machine learning models can learn from new data, which is trained improving their predictive accuracy over time. This approach enhances the accuracy and reliability of predictions by considering a some limitation factors. This project should prioritize ethical considerations, including patient privacy and data security, ensuring that the deployment of machine learning models aligns with ethical standards and regulations. We have very huge dataset and how future machine learning algorithms selected to predict and classify breast cancer and our intention is to reduce error rate with maximum accuracy. This project based on to support a tool for valuable health care professionals by providing additional information and insights to diagnose and treatment. Our Analysis could help you for future analysis and for making easy things and improve the accuracy.

II. LITERATURE SURVEY

The following literature survey provides general description on cause and some works on Breast cancer prediction techniques:

[1]V Chaurisya & S Paul, Wisconsin Breast cancer: The tool used is WEKA and technique used is Statistical Feature Selection. Patient features in the dataset is is sorted out from data materials and statistically tested the data set built on the type of individual data set feature. Then some attributes or some features are tested and selected out , and in each the important feature is selected out after test, and features importance score is calculated. In this XGBoost algorithm is done by repeating ten fold cross validation. The accuracy is 96.45 and the Error rate is 0.33%.

[2] Keles, M.Kaya, Wisconsin Diagnostic Breast cancer dataset: The tool used is Python and the technique used is SVM vs KNN, decision trees and Naives bayes. SVMs map the feature space and input vector of higher dimensionality and this is done to identify the hyper-plane and this separates the data points into the two classes. And the result of this hyper-plane is the marginal distance between the instance and the decision hyper-plane that are closest to the boundary is maximized. The accuracy is up to 96.91% and the error rate is 0.33%.

.

[3] Wang et.al, Wisconsin breast cancer database(1991) and Wisconsin diagnostic breast cancer (1995): The tool used is WEKA and the technique used is PCA. The advantage is that the dimension reduction Wang and Yoon technique , manifests relation to prediction accuracy and efficiency. The accuracy is eight PCs are chosen

[4] Kibeom et.al, Gene expression dataset collection: The tool used is WEAK and the technique used is C4.5,Bagging and ADABOOST Decision trees. It ensemble method helps to combine multiple learners. The accuracy is single C4.5= 95.6%,Bagging C4.5=93.29% and ADABOOST C4.5 =92.62%. and the error rate is sensitivity = 56% and 72%.

III. METHODOLOGY

This study proposes a Machine learning-based approach to detect the breast cancer, similar to algorithm , for the purpose of detecting and predicting disease with high validity and reducing the maximum error rate. The present pattern has been designed on SVM, KNN, GaussianNB and Decision Tree algorithms. In this methodology mainly consists of Data pre- processing, EDA, feature selection and model selection:

1. **Data pre-processing:** The first we collect the data that which is required for our diagnosis and the collected content is checked and used on classification and regression methods. It is a data mining technique that used for converting unmodified data into readable format. The raw data is inconsistent incomplete etc, to resolve this issues the pretreated data is done initially. This is very important step which checks and determines the quality and quantity of data and the model developed. In this case we collect the breast cancer samples which having tumours. This will be training data.

2. **Exploratory Data Analysis:** In this the terms used are visualization of class distribution, scatter plot, histograms, and correlation matrices. In this the dataset is examined to gain or to get the insights into the breast cancer prediction task. The visualized distribution of classes are Malignant and Benign which are use to conlude the cancer or breast tumour. Scatter plot allows us to visualize the relationship between the different features. Histogram gives knowledge on the distribution of each feature individually.

claimed by the screen plot ,which explains 92.6%of total correlation . And ten PCs are selected based on 95% of correlation.

3. **Feature selection:** In Machine learning feature selection also nominated as the variable selection, attribute selection, it is

The process of selection of relevant features for this model construction. Data file and breast cancer feature selection:- data set from Kaggle repository out of many parameters we have selected a very few which are required to diagnose and predict. The important features found are clump-thickness, uniform_cell_size, shape, marginal_adhesion mitosis, class etc.

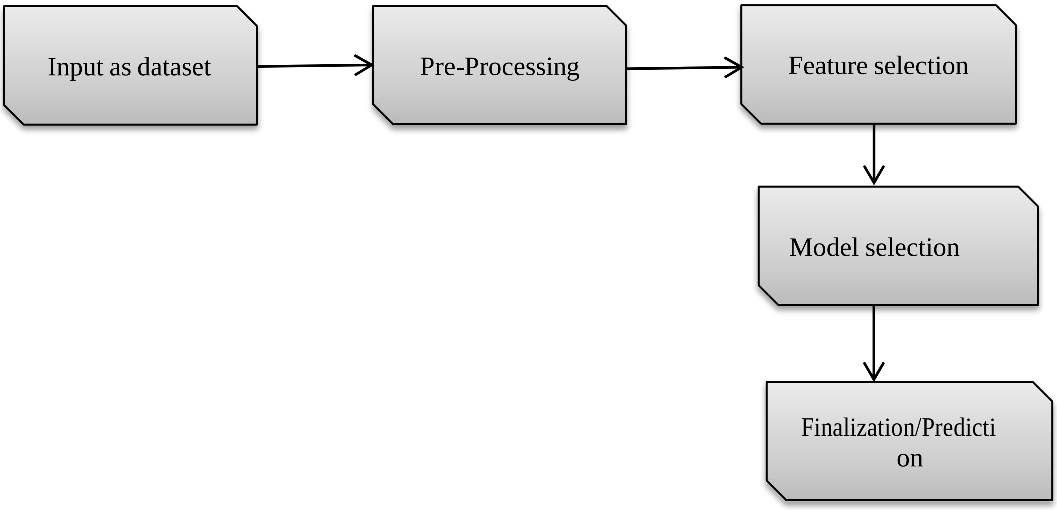


Fig. Methodology

Methods Used:

Support Vector Machine:

This method is powerful classifier chosen to check the optimal hyper-plane to separate classes in a high-dimensional space, best for complex class dimensions. It is also a sub category artificial intelligence algorithm which is used when the number offeatures and number of instances are many. This is very good at pattern recognition problems and used as training algorithm for studying regression and classification. From this the binary classifiers are built through SVM algorithm. Whereeach pointin the SVM model showcased as a referral dots in an n-dimensional space.

Algorithm:

(1) First it find lines that exactly classifies the training dataset.

(2) From this lines it picks which is far from the closest data points.

Decision Tree: Decision tree uses a tree like structure to make decisions based on features, differing interportability And visiability but may be prone to overfiting.

Gaussian Naïve Bayes: NB simple probability algorithm assuming Gaussian distribution effective for text classification and Span filtering with computational efficiency.

KNN: It is non-parametric simple yet effective algorithm for classification and regression in this research. It does by looking „k" nearest or closest data points to a given data point . It does not require any training phase and making it as a popular choices for many applications

Algorithm:

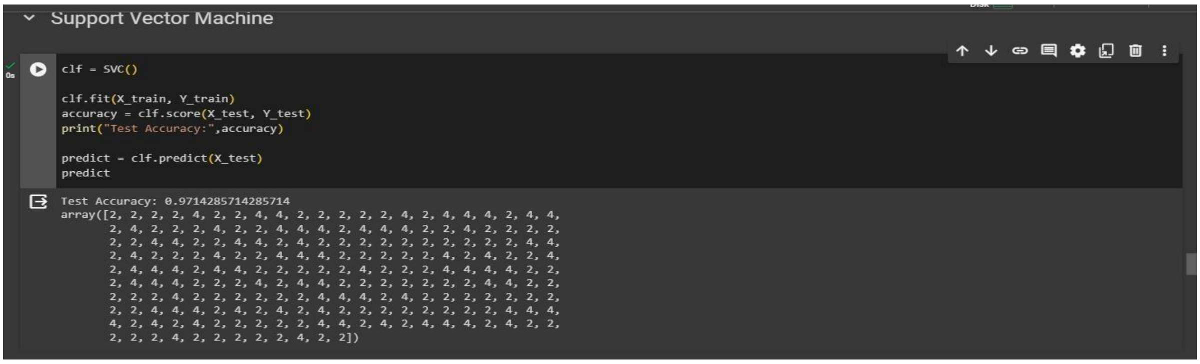
(1) Input the data set.

(2) Split into training and test dataset.

(3) Pick an instance from the testing set and calculate distance with training set.

(4) List distance in ascending order.

(5) Class of instance which as highest common class of the 3 test training (k=3).



Support Vector Machine

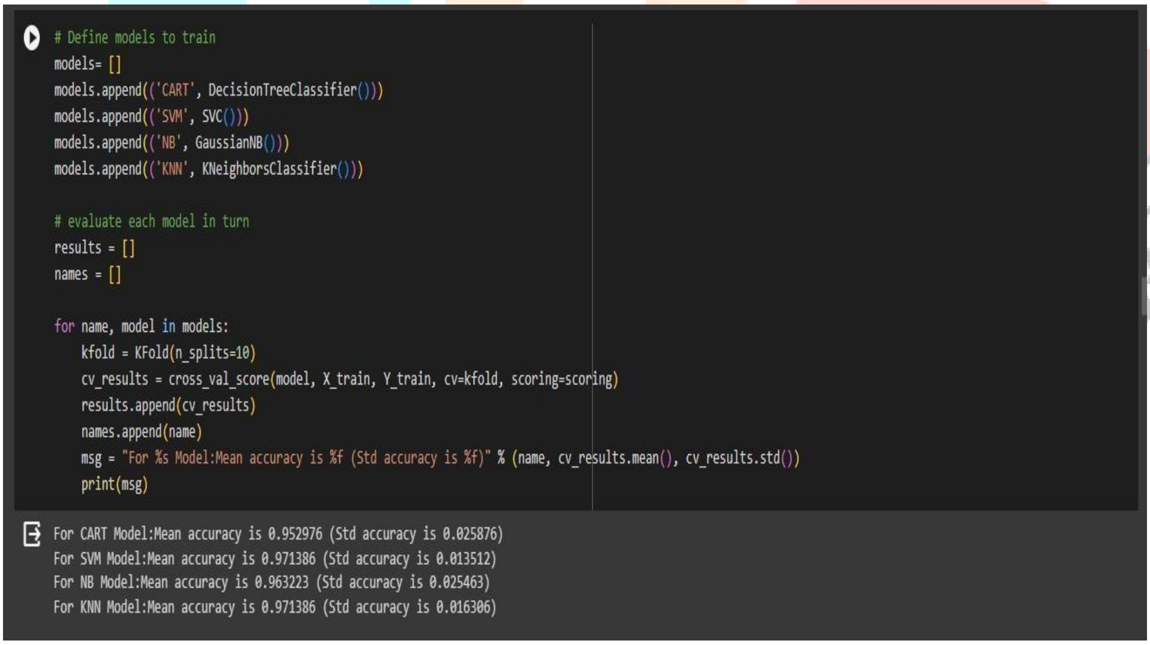
```
clf = SVC()

clf.fit(X_train, Y_train)
accuracy = clf.score(X_test, Y_test)
print("Test Accuracy:",accuracy)

predict = clf.predict(X_test)
predict
```

```
Test Accuracy: 0.9714285714285714
array([2, 2, 2, 2, 4, 2, 2, 4, 4, 2, 2, 2, 2, 2, 4, 2, 4, 4, 4, 2, 4, 4,
       2, 4, 2, 2, 2, 4, 2, 2, 4, 4, 4, 2, 4, 4, 4, 2, 2, 4, 2, 2, 2, 2,
       2, 2, 4, 4, 2, 2, 4, 4, 2, 4, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 4, 4,
       2, 4, 2, 2, 2, 4, 2, 2, 4, 4, 4, 2, 2, 2, 2, 2, 4, 2, 4, 2, 2, 4,
       2, 4, 4, 4, 2, 4, 4, 2, 2, 2, 2, 2, 4, 2, 2, 2, 4, 4, 4, 4, 2, 2,
       2, 4, 4, 4, 2, 2, 2, 4, 2, 4, 4, 2, 2, 2, 2, 2, 2, 2, 4, 4, 2, 2,
       2, 2, 2, 4, 2, 2, 2, 2, 2, 2, 4, 4, 4, 2, 4, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 4, 4, 4, 2, 4, 2, 4, 2, 4, 2, 2, 2, 2, 2, 2, 2, 2, 4, 4, 4,
       4, 2, 4, 2, 4, 2, 2, 2, 2, 2, 4, 4, 2, 4, 2, 4, 4, 4, 2, 4, 2, 2,
       2, 2, 2, 4, 2, 2, 2, 2, 2, 4, 2, 2])
```

I. RESULT

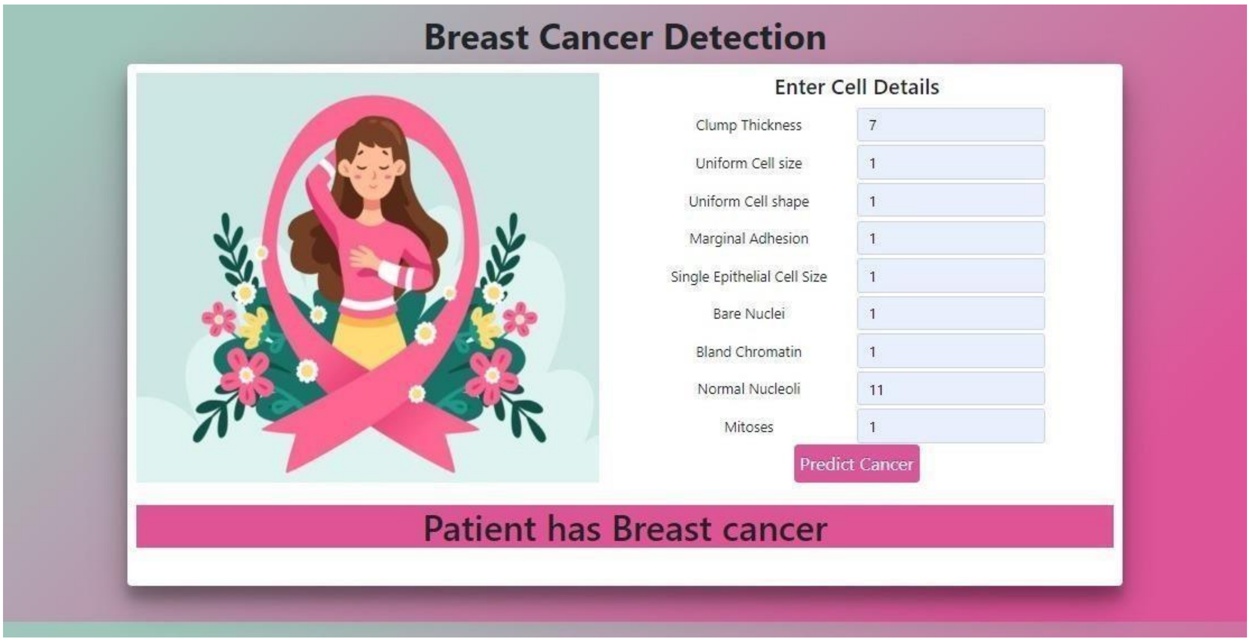
The results obtained from our model‟s is evaluated by it‟s training and testing accuracy.

MODEL	MEAN ACCURACY	STANDARD ACCURACY
NB	**0.963223**	**0.025463**
KNN	**0.971386**	**0.0163**



```
# Define models to train
models= []
models.append(('CART', DecisionTreeClassifier()))
models.append(('SVM', SVC()))
models.append(('NB', GaussianNB()))
models.append(('KNN', KNeighborsClassifier()))

# evaluate each model in turn
results = []
names = []

for name, model in models:
    kfold = KFold(n_splits=10)
    cv_results = cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "For %s Model:Mean accuracy is %f (Std accuracy is %f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)
```

```
For CART Model:Mean accuracy is 0.952976 (Std accuracy is 0.025876)
For SVM Model:Mean accuracy is 0.971386 (Std accuracy is 0.013512)
For NB Model:Mean accuracy is 0.963223 (Std accuracy is 0.025463)
For KNN Model:Mean accuracy is 0.971386 (Std accuracy is 0.016306)
```



Breast Cancer Detection

Enter Cell Details

Clump Thickness: 7
Uniform Cell size: 1
Uniform Cell shape: 1
Marginal Adhesion: 1
Single Epithelial Cell Size: 1
Bare Nuclei: 1
Bland Chromatin: 1
Normal Nucleoli: 11
Mitoses: 1

Patient has Breast cancer

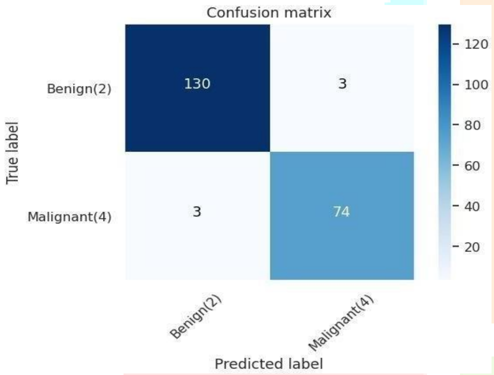


Fig Confusion Matrix.

II. CONCLUSION

In evolution of detection models driven by machine learning offer improved early detection. Intergrating diverse data prioritizingexplainability enhance accuracy. Future efforts should focus on large datasets, real-world evidences and ethical considerations, and need to be explored decentralized models and emerging technologies for more accessible and accurate predictive tools. In thiseach and every method contribute to build a model each results in different accuracies . At last the power ofits methods is shown by effectiveness and efficiency based on accuracy and recall.

III. REFERENCES

[1.] Wang, D. Zhang and Y. H. Huang “Breast cancer prediction using machine learning”(2018), Vol. 66, NO.7.

[2.] Joseph A. Cruz and David S. Wishart “Applications of Machine Learning in cancer prediction and prognosis cancerinformatics”2(3):59-77 February 2007.

[3.] Chen SI, Tseng HT, Hsieh CC. Evaluating the impact of soy compounds on breast cancer using the data mining approach.
. 2020;11(5):4561–70. doi: 10.1039/C9FO00976K.

[4.] Zhang X, Shengli SU, Hongchao WA. Intelligent diagnosis model and method of palpation imaging breast cancer based ondata mining. *Big Data Research* . 2019;5(1):2019005. doi: 10.11959/j.issn.2096-0271.2019005.

[5.] Abdelghani Bellaachia, Erhan Guven, “predicting breast cancer survivability using data mainig techniques”.

[6.] Vikas chaurasia and S.Pal, "using machine learning algorithms for breast cancer risk prediction and diagnosis"(FAMS2016) 83 (2016) 1064-1069.

[7.] Delen, D.; Walker, G.; Kadam, A. Predicting breast cancer survivability: A comparision of three data mining methods.
Artif. Intell. Med. 2005,34,113-127.

[8.] N. Khuriwal, N. Mishra."Mammography Images using Deep Learning Techniques", (2018).

[9]. Preethi S and Aishwaraya Palaniappan, "Brain tumour detection by modified particle swarm optimization algorithm and multi-support vector machine classifier", International Journal of Intelligent Engineering and systems, 2022.

[10]. Preethi and Aishwarya P, "An efficient wavelet-based image fusion for brain tumour detection and segmentation over PET and MRI image", Journal Multimedia tools and applications, doi.org/10.1007/s11042-021-10538-3, 2021.