INTERACTIVE BREAST CANCER PREDICTION USING NAIVE BAYES ALGORITHM

[1]A.T.GAYATHRI, [2]S.T.DEEPA

[1]MPHIL RESEARCH SCHOLAR
DEPARTMENT OF COMPUTER SCIENCE
SHRI SHANKARLAL SUNDARBAI SHASUN JAIN COLLEGE FOR WOMEN, CHENNAI, TAMILNADU,

[2]ASSOCIATE PROFESSOR
DEPARTMENT OF COMPUTER SCIENCE,
SHRI SHANKARLAL SUNDARBAI SHASUN JAIN COLLEGE FOR WOMEN, CHENNAI, TAMILNADU.

Abstract: Breast cancer is most common cancer among women at the present world. The reasons that cause this disease are many and cannot be extracted easily. Further, the process to determine stage of the cancer for a particular patient requires great effort from the doctors. Diagnosis at the early stage is at most important to save the life of the human. Moreover, diagnosing manually takes more amount of time and hence there is a necessity for developing an automated system for early diagnosis. Contribution of Machine learning algorithms for development of such system is lot. In this paper, the objective is to diagnose breast cancer stages of the patient using Naive Bayes algorithm. The dataset used for analysing the performance of the system is taken from UCI repository. This technique helps the physician to take better decision for the breast cancer patient. At last Confusion matrix is used to determine the accurate performance of Naive Bayes Algorithm.

KeyWords: Breast Cancer, Naive Bayes Algorithm, Stages, Healthcare Management, Public Health, Data Mining.

I. INTRODUCTION

In today's world one of the most important that makes it a significant public health problem is Breast Cancer (BC). The oncologist, an expert who diagnose breast cancer by examining the entire medical history, Physical examination of both the breasts and also check for swelling of any lymph nodes, and various image testing like Magnetic Resonance Imaging (MRI) and Ultrasound of breast, X-ray of the breast are performed to determine cancer cells in lymph nodes to confirm metastasis of breast cancer. Based on the necessity, oncologist may also carry out additional tests or procedures.

Alternatively now a days, machine learning algorithms has been widely used for analysis and detection which achieves favourable performance. Our analysis provides a comprehensive guide to sensitivity analysis of model parameters with regard to performance in detection of breast cancer stages. Early diagnosis of breast cancer can boost prediction and survival rate, so that patients can be given a clear idea about clinical treatment at the right time. Using machine learning methods for diagnostic can significantly increase processing speed and on a big scale can make the diagnostic significantly cheaper. The main objective of using naive bayes algorithm is to predictive analytics model to diagnose breast cancer stages of patients considering the attributes like Tumor size, Inv-nodes and Nodecaps. Additionally, the performance of naive bayes algorithm for the given dataset is evaluated by determining accuracy rate, sensitivity, specificity using confusion matrix.

Literature Survey:

Hongchao Song, Aidong Men and Zhuqing Jiang used Empirical mode decomposition (EMD)-based feature extraction method that is more robust to signal misalignment. The statistical features are extracted from the decomposed sub bands of the original signal. The experimental results obtained from clinical data indicate that the detection accuracy is improved by the combination of features from EMD and Principal component analysis (PCA). PCA is one of the most widely used feature extraction methods; however, PCA is negatively impacted by signal misalignment. The estimated malignant-to-normal breast tissue contrast is approximately 2:1 to 10:1 depending on the density of the normal tissue.

Hoda S. Hashemi, Stefanie Fallone and Mathieu Boily Assessed Mechanical Properties of Tissue in Breast Cancer-Related Lymphedema using new novel ultrasound techniques like Ultrasound Elastography. Elastography as ability to identify changes in mechanical properties of the tissue related to detection and staging of lymphedema. Zhiqiong Wang, Mo Li, Huaxia Wang determined Breast Cancer Detection Using Extreme Learning Machine Based on Feature Fusion with CNN Deep Features. The main idea is to apply deep features extracted from CNN to the two stages of mass detection and mass diagnosis. In the stage of mass detection, a method based on sub-domain CNN deep features and US-ELM clustering is developed. In the stage of mass diagnosis, an ELM classifier is utilized to classify the benign and malignant breast masses using a fused feature set, fusing deep features, morphological features, texture features, and density features. Ravi K. Samala, Heang-Ping Chan and Lubomir Hadjiiski used Digital Breast Tomosynthesis to diagnose Breast Cancer, Multi-Stage Transfer Learning with Deep Neural Nets is the methodology used to determine effects of Training Sample Size. Dong Wei, Susan Weinstein, Meng-Kang Hsieh used the concept of Segmentation of Breast Three-Dimensionally in Sagittal and Axial Breast MRI to determine Chest-Wall Line Detection using Field Modelling. Dongdong Sun, Minghui Wang performed research work on human breast cancer prognosis prediction using neural network by collaborating multiple data. S. Dencks, M. Piepenbrock and T. Opacic developed Clinical Pilot Application of Super-resolution using US Imaging in Breast Cancer. Shailima Rampogu, Ayoung Baek, Rohit Bavi performed Identification of Novel Scaffolds with Dual Role as Antiepileptic and Anti-Breast Cancer

II. EXISTING SYSTEM

In the existing system the patterns frequently appearing in the Tumors with the same label can be regarded as a potential diagnostic rule. Subsequently, the diagnostic rules are utilized to construct component classifiers of the Adaboost algorithm. The AdaBoost learning is performed to discover effective combinations and integrate them into a strong classifier. The proposed approach has been validated using a large ultrasounic dataset of 1062 breast tumor instances (including 418 benign cases and 644 malignant cases) and its performance was compared with several conventional approaches. The experimental results show that the proposed method yielded the best prediction performance, indicating a good potential in clinical applications.

Boosting-general method of converting rough rules of thumb into highly accurate predication rule. For a given sufficient data, a boosting algorithm can provably construct single classifier with very high accuracy and studied some advantages of boosting

Advantages

1. Less error based on ensemble method.
2. Suitable if the initial model is pretty bad.

Drawbacks:

1. Its cannot work on top features and find-out the accuracy, Recall, Precision, Confusion matrix and compare it with our old result.
2. It’s cannot work on using the popular machine learning algorithm to find out the features importance.

III. PROPOSED SYSTEM

Following are the steps performed to determine the breast cancer stage of a patient

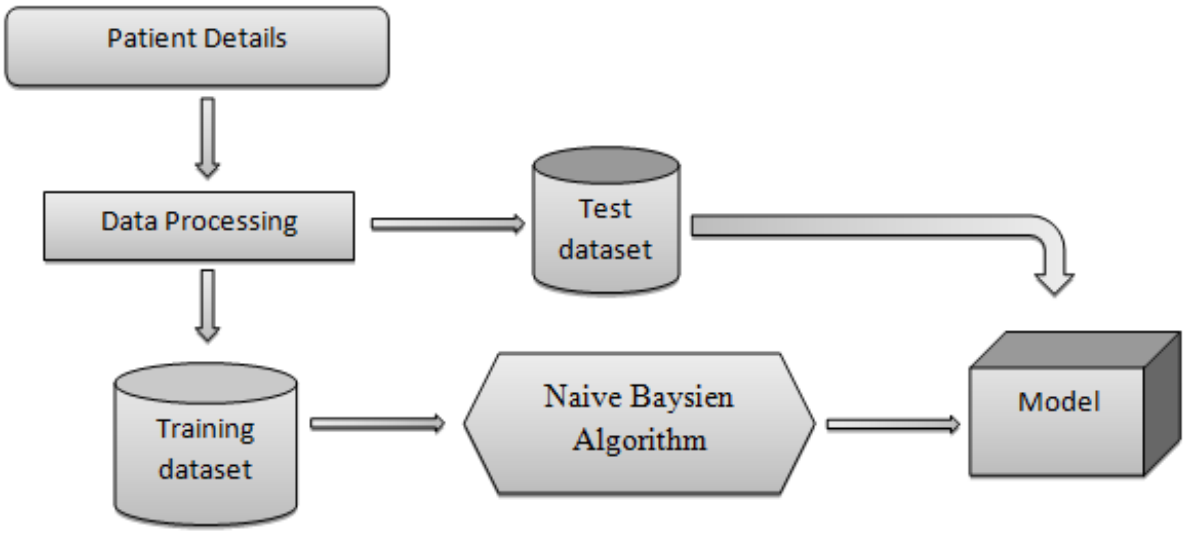

Fig: Architecture of Proposed model

IV. SOFTWARE DESCRIPTION

Anaconda Navigator

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows users to launch applications and manage conda packages, environments and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository, install them in an environment, run the packages and update them. It is available for Windows, macOS and Linux.

The following applications are available by default in Navigator:

- JupyterLab
- Jupyter Notebook
- QtConsole
- Spyder
- Glueviz
- Orange
- Rstudio
- Visual Studio Code

V. METHODOLOGY

Breast Cancer Dataset:

The data set used in study is taken from UCI repository. The data set has 10 attributes and 287 rows and contains following variables

Table 1: Data set Description

Variable	Description
Age	Age of patient (At the time of diagnosis)
Menopause	Menopause status of the patient
Tumor-size	Patient tumor size
Inv-nodes	Range of axillary lymph nodes showing breast cancer at the time of historical examination
Nodecaps	Penetration of the tumor in the lymph node capsule or not
Deg-malig	Grade of tumor
Breast	Position of breast cancer
Breast-quad	If the nipple consider as a central point the breast may be divided in to four quadrants
Irrad	Irradiation: Patient radiation therapy
Class	Depends on reappearing symptoms of breast cancer in the patients after treatment

Preprocessing

To obtain the best result out of the analysis the dataset need to be preprocessed properly. The data so collected might contain some missing values that may lead to inconsistency. To gain better results data need to be preprocessed so as to improve the efficiency of the algorithm. The outliers have to be removed and also variable conversion need to be done. Based on the correlation among the attributes the most important and significant attributes are Tumor-size, Inv-Node, Node caps to determine the grade level of the cells infected and to determine TNM level

Grade:

Grading is methodology used in the TNM staging system wherein it indicates the cell appearance, growth and the speed at which the cancer spread to the other parts. The cells are graded in three categories 1, 2, and 3. One of the types of cancer Ductal Carcinoma in Situ (DCIS) follows different grading system as such low, medium and high instead of 1, 2, and 3.

Grade 1 – In this category the size of the cells looks small and uniform. The growth of the cell is also slow when compared to other grades.

Grade 2 – In this category the size of the cell is little bigger and different in shape than normal cells. The growth is faster in this category.

Grade 3 – In this category the cells looks uniform to normal cells and even the growth is faster.

TNM staging system:

The most commonly used tool that doctors uses to describe the stage is the standard staging system TNM system.

- **Tumor (T):** How large is the primary tumor? Where is it located?
- **Node (N):** Has the tumor spread to the lymph nodes? If so, where and how many?
- **Metastasis (M):** Has the cancer spread to other parts of the body? If so, where and how much?

The **T category** is further sub-divided into T0, Tis, T1, T2, T3, or T4 categories depending on the size of the tumor. Higher the T category implies larger the tumor and wider the spread to the tissues near the breast. Tis in T category imply carcinoma in situ. The **N category** is further sub-divided into N0, N1, N2, and N3 categories which indicates how far the cancer has spread to lymph nodes or it indicates number of lymph nodes affected. N is higher when more lymph nodes are affected. The **M category** is further sub-divided into M0 and M1 which represents lab test reports but it is not part of the pathology report generated after breast cancer surgery. The result of T, N, and M categories is used to determine, the overall stage of the cancer.

Work flow diagram

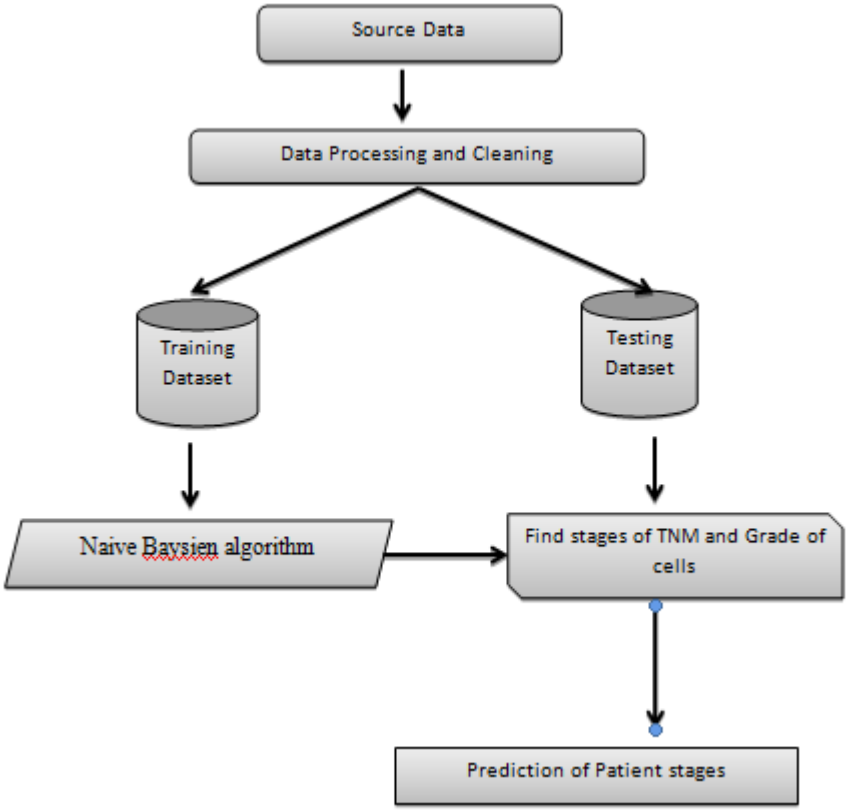

Fig: Workflow Diagram

VI. CLASSIFICATION METHODS

Naive Bayes Classifier:

In the world of Data science one among the classification algorithm is the Naive Bayes Algorithm which is based on the Bayes's Theorem. The Naive Bayes is a popular method for creation of statistical predictive models. This algorithm is most commonly used for problems related to prediction for ease of implementation and usage. This classification algorithm is used to analyse the relationship between the attribute and various classes to determine the prediction or probability of the problem.

According to this study for the given patient conditions, each tuple classifies the conditions as fit ("Recurrence event") or unfit("N0-Recurrence event") for class.

The dataset is classified into two categories, such as **feature matrix** and the **response vector**.

- Feature matrix represents the vector (rows) of the Dataset and each vector contains values which are **Dependent**. Based on the dataset considered for the study, features are 'age', 'menopause', 'tumor-size', 'inv-nodes', 'node-caps', deg-malig', 'breast', 'breast-quad' and 'irradiat'.

- Response vector represents actual value of the variable such that the each row in the Feature Matrix. Similarly, the name of the variable in class is the 'Class' itself.

Assumption:

Naive Bayes algorithm assumes that each feature in the dataset is Independent and Equal contribution to the result.

Based on the relationship between features in dataset, the assumptions can be understood as:

- Each feature is independent of each other.
- Secondly, Equal importance is given to each Feature.

Baye's Theorem:

In terms of Theoretical probability and event, given the probability of another event the Bayes Theorem determines the probability for occurring an event. Mathematical representation of Baye's Theorem

$P(A/B) = (P(B/A)\ P(A)) / P(B)$

Where A and B are events and $P(B) \neq 0$

P (A/B): Possibility of event A occurrence given that event B is True.

P (B/A): Possibility of event B occurrence given that event A is True.

P (A) and P (B) are the actual possibilities that occur, which are independent of each other.

Now, according to the given dataset Bayes' theorem can be applied in following way:

$P(Y/N) = (P(N/Y)\ P(Y)) / P(N)$

Where, Y is variable class and N is feature vector which is dependent of size *n* such that:

N = (n1, n2, n3.....)

Basically according to the study, P (N|Y) represents, the probability of "Not class" given the patient conditions attributes are "tnm", "stage" and "age".

Generally in Gaussian Naive Bayes, values that continuous are associated with each feature and are assumed to be distributed hence it is called **Gaussian distribution**. A Gaussian distribution is also named as Normal distribution.

For example: implementation of Gaussian Naive Bayes classifier using scikit-learn

```
# load the iris dataset
Import pandas as p
iris = p.read_csv("dataset.csv")

# storing the feature matrix (N) and response vector (Y)
N = iris.data
Y = iris.target

# splitting N and Y as training and testing datasets
from sklearn.model_selection import train_test_split
N_train, N_test, Y_train, Y_test = train_test_split(N, Y, test_size=0.4, random_state=1)

# model is trained using training dataset
from sklearn.naive_bayes import GaussianNB
gnb = GaussianNB()
gnb.fit(N_train, Y_train)

#prediction made on testing data
Y_pred = gnb.predict(N_test)

# comparing actual response values (Y_test) with predicted response values (Y_pred)
Y from sklearn import metrics
print("Gaussian Naive Bayes model accuracy(in %):", metrics.accuracy_score(y_test, y_pred)*100)
```

The performance of the Classifier is validated by determining the Classification accuracy in terms of Sensitivity and Specificity. And F1 score is also determined to validate the performance.**F1 Score** is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

The sensitivity or the true positive rate (TPR) is defined by TP / (TP + FN); while the specificity or the true negative rate (TNR) is defined by TN / (TN + FP); the accuracy is defined by (TP + TN) / (TP + FP + TN + FN); and the F1 score is defined by 2*(Recall * Precision) / (Recall + Precision).

True positive (TP) = number of positive samples correctly predicted.
True negative (TN) = number of negative samples correctly predicted.
False negative (FN) = number of positive samples wrongly predicted.
False positive (FP) = number of negative samples wrongly predicted as positive.

The values those are determined are represented in the form of confusion matrix

Table 2: Confusion Matrix

Actual	**Predicted**		**Total**
	Negative	**Positive**	
Negative	**59**	**0**	**59**
Positive	**0**	**25**	**25**
Total	**59**	**25**	**84**

Table 3: Performance of Test data

Method	Accuracy	Specificity	Sensitivity	F1 Score	Recall	Precision
Naive Bayes	100	100	100	100	100	100

VII. CONCLUSION

GUI Based Predicting Breast Cancer Using Naive Bayes Algorithm proposed to determine the breast cancer stage for a patient and to determine the grading level of a cell as show to be accurate in the performance. The performance of the Naive Bayes algorithm is evaluated by determining Sensitivity and Specificity using confusion matrix. Therefore, Naive Bayes Classifier best suits for diagnosis of Breast Cancer.

REFERENCES

1. S. Dencks, *Member, IEEE,* M. Piepenbrock, T. Opacic, B. Krauspe, E. Stickeler, F. Kiessling, G. Schmitz, *Senior Member, IEEE "*Clinical Pilot Application of Super-resolution US Imaging in Breast Cancer".
2. Dongdong Sun, Minghui Wang and Ao Li "A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data".
3. Hoda S. Hashemi, Stefanie Fallone, Mathieu Boily, Anna Towers, Robert D. Kilgour,and Hassan Rivaz "Assessment of Mechanical Properties of Tissuein Breast Cancer-Related Lymphedema Using Ultrasound Elastography".
4. Mr. D R Umesh, Thilak C R " Predicting Breast Cancer Survivability Using Naïve Bayes and C5.0 Algorithm", International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online) Vol. 3, Issue 2, pp: (802-807), Month: April - June 2015.
5. Amir H. Golnabi*, *Member, IEEE,* Paul M. Meaney, *Fellow, IEEE,* Shireen D. Geimer, and Keith D. Paulsen, *Fellow, IEEE "*3D Microwave Tomography Using the Soft Prior Regularization Technique: Evaluation in Anatomically-Realistic MRI-Derived Numerical Breast Phantoms".
6. Chen Peng, Yang Zheng, and De-Shuang Huang*, *Senior Member, IEEE "*Capsule Network based Modeling of Multi-omics Data for Discovery of Breast Cancer-related Genes".
7. ZHIQIONG WANG, MO LI, HUAXIA WANG, HANYU JIANG, YUDONG YAO(Fellow, IEEE), HAO ZHANG, AND JUNCHANG XIN," Breast Cancer Detection Using Extreme Learning Machine Based on Feature Fusion with CNN Deep Features".
8. Ravi K. Samala, Heang-Ping Chan, Lubomir Hadjiiski, Mark A. Helvie, Caleb D. Richter, Kenny H. Cha "Breast Cancer Diagnosis in Digital Breast Tomosynthesis: Effects of Training Sample Size on Multi-Stage Transfer Learning using Deep Neural Nets".
9. Chen Peng, Yang Zheng, and De-Shuang Huang*, *Senior Member, IEEE "*Capsule Network based Modeling of Multi-omics Data for Discovery of Breast Cancer-related Genes".
10. S. Dencks, *Member, IEEE,* M. Piepenbrock, T. Opacic, B. Krauspe, E. Stickeler, F. Kiessling, G. Schmitz, *Senior Member, IEEE "*Clinical Pilot Application of Super-resolution US Imaging in Breast Cancer".
11. Amritpal Singh, Ryerson University, Toronto, John Dillon Odette Cancer Centre, Toronto, Ananth Ravi, Odette Cancer Centre, Toronto "Construction and Characterization of a Novel Single Pixel Beta Detector for Intra-operative Guidance in Breastconserving Surgery".
12. Yangqin Feng, Lei Zhang, Senior Member, IEEE, Juan Mo "Deep Manifold Preserving Autoencoder for Classifying Breast Cancer Histopathological Images".
13. Hang Song, Shinsuke Sasada, Norio Masumoto, Takayuki Kadoya, Noriyuki Shiroma, Makoto Orita, Koji Arihiro, Morihito Okada and Takamaro Kikkawa, Fellow, IEEE "Detectability of Breast Tumors in Excised Breast Tissues of Total Mastectomy by IR-UWB-Radar- Based Breast Cancer Detector".
14. Huangjing Lin, Student Member, IEEE, Hao Chen*, Member, IEEE, Simon Graham, Student Member, IEEE, Qi Dou, Member, IEEE, Nasir Rajpoot, Senior Member, IEEE, and Pheng-Ann Heng, Senior Member, IEEE " Fast ScanNet: Fast and Dense Analysis of Multi-Gigapixel Whole-Slide Images for Cancer Metastasis Detection".
15. Yangqin Feng, Lei Zhang, Senior Member, IEEE, Juan Mo "Deep Manifold Preserving Autoencoder for Classifying Breast Cancer Histopathological Images".
16. Firat Ismailoglu, Rachel Cavill, Evgueni Smirnov, Shuang Zhou, Pieter Collins, Ralf Peeters "Heterogeneous Domain Adaptation for IHC Classification of Breast Cancer Subtypes".

17. Lin Yuan, Le-Hang Guo, Chang-An Yuan, You-Hua Zhang, Kyungsook Han, Asoke K. Nandi, Barry Honig, and De-Shuang Huang "Integration of Multi-omics Data for Gene Regulatory Network Inference and Application to Breast Cancer".
18. Seung Yeon Shin, Soochahn Lee_, Member, IEEE, Il Dong Yun, Member, IEEE, Sun Mi Kim, and Kyoung Mu Lee, Senior Member, IEEE "Joint Weakly and Semi-Supervised Deep Learning for Localization and Classification of Masses in Breast Ultrasound Images".
19. Qi Qi, Yanlong Li, Jitian Wang, Han Zheng, Yue Huang, Xinghao Ding, Gustavo Kunde Rohde "Label-efficient Breast Cancer Histopathological Image Classification".
20. Eleftherios Kontopodis*, Maria Venianaki*, Georgios C. Manikis, Katerina Nikiforaki, Ovidio Salvetti, Efrosini Papadaki, Georgios Z. Papadakis, Apostolos H. Karantanas, Kostas Marias " .