Tumour Stage Classification Using Machine Learning

line 1: 1st Given Name Surname

line 2: dept. name of organization (of Affiliation)

line 3: name of organization (of Affiliation)

line 4: City, Country

line 5: email address or ORCID

Latika Adhikari

Electronics And Communication Engineering-Artificial Intelligence Department

IGDTUW

New Delhi,India

[latika064bteceai23@igdtuw.ac.in](mailto:latika064bteceai23@igdtuw.ac.in)

Madhu

Electronics And Communication Engineering-Artificial Intelligence Department

IGDTUW

New Delhi,India

[madhu066bteceai23@igdtuw.ac.in](mailto:madhu066bteceai23@igdtuw.ac.in)

Jannat

Electronics And Communication Engineering-Artificial Intelligence Department

IGDTUW

New Delhi,India

[jannat044bteceai23@igdtuw.ac.in](mailto:jannat044bteceai23@igdtuw.ac.in)

*Abstract*—Breast Tumor is one of the most common types of tumors affecting women all around the world, early and accurate detection of tumor is necessary for effective treatment. This research paper focuses mainly on classifying stages of breast tumor by the advanced use of machine learning models. Several machine learning models that work on multiple classification are used such as Logistic Regression, Decision Tree, Support Vector Machine, Naive Bayes, K-Nearest Neighbors and XGBoost. Our study signifies the potential of machine learning in accurately classifying the breast tumor.

Keywords— Breast Tumor, multiple-classification, machine learning, machine learning algorithms.

1. Introduction

Breast tumor is the one of the most common cancers among women worldwide, with an estimated 2.3 million new cases diagnosed in 2020 [1]. Early detection and accurate prediction of stage of tumor is very important for improving treatment planning. Machine Learning algorithms plays a vital role in model development for predicting the tumor stage. These algorithms make use of large dataset to train the model and being capable of differentiating between different stages of tumor. The main focus of this paper is on using different machine learning algorithms to train the dataset and use the most accurate algorithm to create the multi-classification model. The findings of this research have remarkable implication for improving breast cancer diagnosis, treatment planning, and finally resulting in increasing survival rate of the patients.

II. Literature Review

In research, the classification of Breast cancer stages has been crucial research, particularly with the use of machine learning techniques [2].

In previous research, it has shown text and data mining methods are becoming important in the cancer precise prediction and classification of breast cancer stages.

There are numerous machine learning models, and their influence on predicting cancer therapy response is explored. Physicians and scientists have several techniques for the identification of cancer, which includes genetic analysis, symptoms-based analysis, early phase screening, etc. Are used to diagnose and increase the survival outcomes of breast cancer patients but not metastatic spread [8].

Hence in this research, we are developing a model that classify the tumor stages and shows how crucial is to informing treatment strategies using machine learning algorithms and techniques that advances the system, reduces human errors and lowers manual mistakes.

The extensive understanding of how various factors influence tumor stages can contribute to the development of personalized medicine approaches, where treatment is specialized based on individuals' cancer stages classification and profile.

Therefore, the early detection of cancer stage of breast cancer is crucial [10].

***III. Dataset Description***

**DATASET SOURCE-**

1. **DATASET FEATURE- We are working on a breast cancer data set that has many like tumor stages. Its columns Help us to find out classification over breast cancer.**
2. Patient\_ID- This is the unique feature as its unique from person to person.
3. Age- Age plays a significant role in detecting the cancer as its treatment reponse can vary .
4. Gender- Gender helps to find out the cancer risk ,tumor stage and its treatment.Breast cancer is identified in women mostlt and men can also develop.
5. Protein1, Protien2, Protein3, Protein4 - Protein act as a biomakers,helps to understand tumor behavior and its treatment. These columns provide quantitative measurements that involved in tumour biology.
6. Tumour\_Stage- It is a categorical variable ,represent the tumor stage .Plays a key role i deciding the treatment plans .
7. Histology- Histology involves analysis of tumor tissue under a microscope.Type of breast cancer affect treatment test and its outcome .
8. ER Status- ER cells tell whether tumor cell have estrogen receptors or not.
9. PR Status- PR cells tell whether tumor cell have progesterone receptors in the tumor cell or not.
10. HER2 Status- HER2 Status shows the presence of the epidermal growth factor receptors 2 protein, which is significant for determining the most effective treatment approach [9].
11. **DATA PROCESSING-**

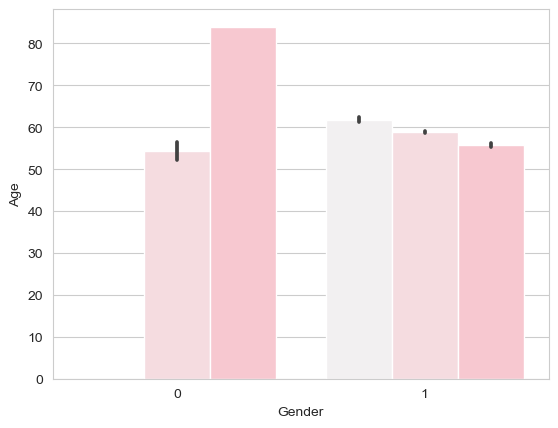
1. Data Cleaning: Clean the data to make sure there is no error and inconsistencies in data:

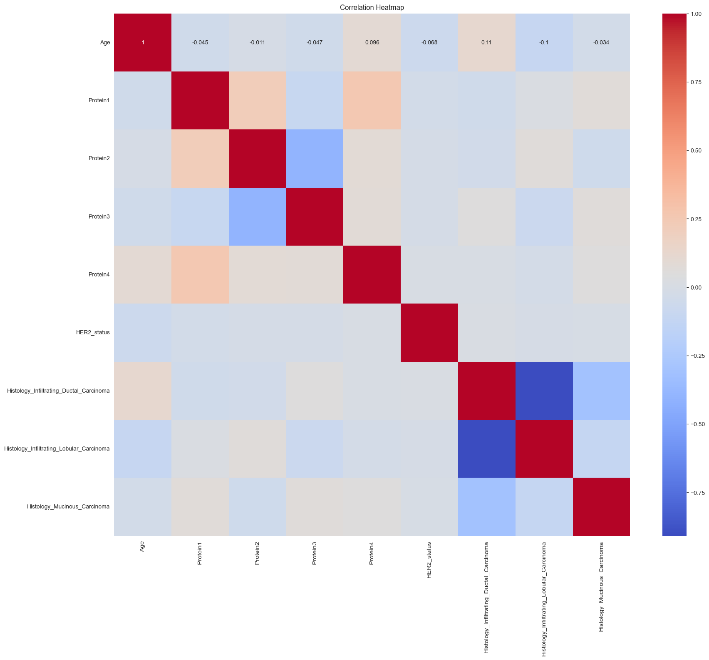
* Handling Missing Data: identify the null values is data and drop all the null values to make the data consistent.
* Removing duplicates: removing the duplicate data to biasing of the result.
* Dropping unwanted columns: dropping the columns that are not useful in training the model to make model more accurate.
* Standardizing data: Standardizing data to make sure that contribution of each feature is on same scale.

2. Data Transformation: Converting data to make it suitable for analysis:

* Datatype conversion: Converting object type data to float and integer type for better analysis of data.
* Creating dummies: Creating dummies of ‘Histology’ column and converting it into integer type data.
* Encoding Categorical Variables: Converting ‘Tumor stage’ into numerical format for analysis.

3. Data Visualization: Creating graphs for better understanding of relations between the different columns of data:

* Barplot: creating a barplot with Gender at x-axis and Age at y-axis. It shows the patient of which gender and at what age suffers from which stage of tumor.
* 
* Distribution graph: creating a normal distribution graph of age and its frequency.
* Correlation table and heatmap: To find the correlation among different columns of data.



4. Model Implementation: Implementing different algorithms to find best fit for model:

* Feature selection: Selecting the best 5 features for model analysis.
* Splitting data: Splitting the dataset into training data and testing data.
* Training data: Training the data on each machine learning algorithm.
* Evaluating classifiers: finding accuracy of training and testing data by different classifiers and making confusion matrix to analyze the best among all.
* Finding the best classifier to proceed the model.

5. Deploying Model: making an interface of model to take inputs and provide the desired output:

* Gradio installation: installing gradio to make the interface of our model.
* Creating interface: creating interface of model to take inputs such as age, her2 status, histology infiltrating ductal carcinoma, histology infiltrating lobular carcinoma and histology mucinous carcinoma and providing the predicted stage of tumor on the basis of analysis done by model.

***IV.*** Methodology

* Breast cancer staging show a multi-class classification Test Because there are different Breast cancer and tumor stage. This model predicts and classify the exact stage of the tumor.
* By classifying Tumor i different stages, the model offers more detailed insights, facilitating in the customization of treatment and enhancing patient care based on the stage (Tumor)
* Factors like age, gender, protein levels and histology play a crucial role in determining tumor stages. As multiclass classification approach utilizes this data and make more precise prediction.

Methodology used in our Breast Cancer:

## 1. Logistic Regression

In the context of Tumor classification Logistic Regression

Evaluate the chance of a tumor being in a specific stage based

On input features. It is particularly important for assessing the impact of each feature on the predicted outcome [6].

## 2. Decision Tree

A Decision Tree is a flowchart model, where each branch corresponds to a decision rule and each leaf represents an outcome. Decision Tree can work with both numerical and categorial data. It suits diverse dataset also.

3.Random Forest [5]

Random Forest creates multiple decision trees and combine all the prediction to enhance the accuracy.

By Applying the strengths of multiple Trees, random forest improves predictive performance and is Resistant to noise in the dataset.

4. Support Vector Machine (SVM)

Description: A classification algorithm that finds the optimal hyperplane to separate different classes in a high-dimensional spaces.SVM is effective for complex datasets and can handle non-linear relationships, potentially improving classification accuracy.

5. Naive Bayes

A probabilistic classifier based on Bayes' theorem, assuming independence among features. It serves as a fast and efficient baseline model, particularly useful for initial evaluations of classification performance.

## 6. K-Nearest Neighbors (KNN)

The k-nearest neighbors (KNN) algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual point. It is one of the popular classifiers used in Machine learning [7].

7. XGBoost, short for extreme gradient boosting, is a Boosting algorithm [3]. Both XGBoost and random forest are integration algorithms based on the decision tree. Different from the Bagging algorithm, boosting algorithm builds weak learners one by one, accumulating multiple weak learners through continuous iteration [4].

For each classifier, the following steps were conducted:

* **Model Training:** Classifiers were trained using the training dataset (X\_train, y\_train).
* **Predictions:** Predictions were made on both the training set and test set (X\_test).
* **Confusion Matrix:** A confusion matrix was generated for test set predictions to evaluate classification performance.
* **Accuracy Calculation:** Accuracy was calculated for both training and test datasets to assess model generalization.
* **Classification Report:** A classification report including precision, recall, and F1-score was generated for each tumor stage to provide a detailed performance evaluation.

## For evaluating model performance

**Accuracy**: This metric measures the overall correctness of the model by calculating the ratio of correctly predicted instances to the total instances. It is useful for understanding the general effectiveness of the model, particularly when the classes are balanced.

***V.Implementation***

1. Software and Tools

* Programming Language: Python
* Libraries:
  + **Scikit-learn**: Model training, evaluation, and feature selection
  + **Pandas and NumPy**: Data manipulation and computation
  + **Matplotlib and Seaborn**: Visualization of data, confusion matrices, and ROC curves
  + **Jupyter Notebooks/Google Collab**: Code execution and interactive analysis

#### **Model Training**

**Process**: Models were trained on labeled datasets, focusing on minimizing overfitting and ensuring generalization.

* **Feature Selection**: Relevant features were selected
* Comparing Different Models

#### **VI. Results Visualizations:**

All models show promising results in classifying the stage of breast tumor.

1. **LOGISTIC REGRESSION**

Confusion Matrix for Logistic Regression:

[[ 0 331 7]  
 [ 0 1046 29]  
 [ 0 458 29]]

Logistic Regression Accuracy of training data: 56.44%

Logistic Regression Accuracy of testing data: 56.58%

precision recall f1-score support  
1 0.00 0.00 0.00 338  
2 0.57 0.97 0.72 1075  
3 0.45 0.06 0.11 487  
  
accuracy 0.57 1900  
macro avg 0.34 0.34 0.27 1900  
weighted avg 0.44 0.57 0.43 1900  
  
**B) DECISION TREE**

Confusion Matrix for Decision Tree:

[[160 144 34]  
 [106 899 70]  
 [ 50 192 245]]

Decision Tree Accuracy of training data: 71.95%

Decision Tree Accuracy of testing data: 68.63%

precision recall f1-score support  
1 0.51 0.47 0.49 338  
2 0.73 0.84 0.78 1075  
3 0.70 0.50 0.59 487  
accuracy 0.69 1900  
macro avg 0.65 0.60 0.62 1900  
weighted avg 0.68 0.69 0.68 1900

**C) RANDOM FOREST**

Confusion Matrix for Random Forest:

* [[149 152 37]  
   [ 94 919 62]  
   [ 40 208 239]]

Random Forest Accuracy of training data: 71.92%  
Random Forest Accuracy of testing data: 68.79%

precision recall f1-score support  
1 0.53 0.44 0.48 338  
2 0.72 0.85 0.78 1075  
3 0.71 0.49 0.58 487  
  
accuracy 0.69 1900  
macro avg 0.65 0.60 0.61 1900  
weighted avg 0.68 0.69 0.68 1900

**D) SUPPORT VECTOR MACHINE**

Confusion Matrix for Support Vector Machine:   
 [[ 5 333 0]  
 [ 0 1066 9]  
 [ 0 435 52]]

Support Vector Machine Accuracy of training data: 59.75%

Support Vector Machine Accuracy of testing data: 59.11%

precision recall f1-score support  
1 1.00 0.01 0.03 338  
2 0.58 0.99 0.73 1075  
3 0.85 0.11 0.19 487  
accuracy 0.59 1900  
macro avg 0.81 0.37 0.32 1900  
weighted avg 0.73 0.59 0.47 1900

**E) NAÏVE BAYES**

[[ 5 9 324]  
 [ 0 59 1016]  
 [ 0 0 487]]  
Naive Bayes Accuracy of training data: 27.05%  
Naive Bayes Accuracy of testing data: 29.00%

precision recall f1-score support  
1 1.00 0.01 0.03 338  
2 0.87 0.05 0.10 1075  
3 0.27 1.00 0.42 487  
accuracy 0.29 1900  
macro avg 0.71 0.36 0.18 1900  
weighted avg 0.74 0.29 0.17 1900

F)  **K-NEAREST NEIGHBORS**

Confusion Matrix for K-Nearest Neighbors:   
 [[146 165 27]  
 [140 862 73]  
 [ 44 191 252]]  
K-Nearest Neighbors Accuracy of training data: 68.58%  
K-Nearest Neighbors Accuracy of testing data: 66.32%  
 precision recall f1-score support  
1 0.44 0.43 0.44 338  
2 0.71 0.80 0.75 1075  
3 0.72 0.52 0.60 487  
accuracy 0.66 1900  
macro avg 0.62 0.58 0.60 1900  
weighted avg 0.66 0.66 0.66 1900

**COMPARISION OF ALL MODELS -** All models showed promising results in classifying the stages of tumour, with XGBoost achieving the highest accuracy of 68.632% across all tumour stages.

|  |  |  |  |
| --- | --- | --- | --- |
|  | ML Model | Training Accuracy | Testing Accuracy |
| 0 | Logistic Regression | 56.440 | 55.580 |
| 1 | Decision Tree | 71.950 | 68.630 |
| 2 | Random Forest | 71.940 | 63.680 |
| 3 | Support Vector Machine | 59.750 | 59.110 |
| 4 | Naive Bayes | 27.050 | 29.000 |
| 5 | K-Nearest Neighbors | 66.500 | 55.160 |
| 6 | XGBoost Model | 71.949 | 68.632 |

# ***VII. Discussion***

Highlight that the Random Forest Models achieved the high training accuracy (71.940%), but there was a noticeable drop in test accuracy (63.680%), indicating potential overfitting

In our model performances align with what is commonly observed in the literature, with K-Nearest Neighbors and Random Forests showing typical signs of overfitting, XGBoost and Decision Tree demonstrating strong performance, and Naive Bayes underperforming. These insights validate your model's results in the context of existing research.

* **Alignment with Expectations**: Our model’s performance is consistent with existing literature, showing expected overfitting in Naive Bayes and Random Forests, and strong results from XGBoost.
* **Reliability of Findings**: Our findings are credible and reliable because they match what has been reported in similar studies. This helps in reinforcing the validity of your results and conclusions.

**Limitations**: Data quality issues, potential class imbalance, and the exclusion of advanced models may limit the generalizability of the findings.

# ***VIII. Conclusion***

#### Summary

**Model Performance**:

* **Top Performers**: Random Forest model achieved the highest training accuracy (71.95%) but showed reduced test accuracy (63.68%) indicating potential overfitting.
* **Balanced Models**: XGBoost and Decision Tree demonstrated strong performance with good training accuracy (71.95%) and decent test accuracy (68.632% and 68.630% respectively), suggesting effective generalization.
* **Underperformers**: Naive Bayes had the lowest accuracy, both in training (27.05%) and testing (29.00%), indicating it is less suited for this dataset.

**Observation**: The significant drop in test accuracy for Decision Trees and Random Forests suggests overfitting, where models perform well on training data but poorly on testing data.

**Generalization**:

* **Effective Models**: XGBoost and Decision Tree showed better generalization compared to other models, making them more reliable for predicting tumor classifications.

**Model Suitability**:

* **Best Models**: XGBoost and Decision Tree are recommended for further development due to their balanced performance and robustness.

**Future Directions**:

* **Data Augmentation**: If data is limited, use augmentation techniques to increase the diversity of your training set.
* **Handling Class Imbalance**: Apply methods such as SMOTE (Synthetic Minority Over-sampling Technique) or class weighting to address any class imbalance issues.
* focus on improving generalization through regularization, cross-validation, and exploring additional models like neural networks or ensemble methods.

#### 

#### 

#### **IX. References**

1. Shreshtha Malvia et al.,2017 "Epidemology of breast cancer in Indian Women: Breast Cancer epidemiology", <https://onlinelibrary.wiley.com/doi/10.1111/ajco.12661> [Google Scholar ].
2. ‘WHO Breast cancer’, *WHO*.

<http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/>

(accessed Feb. 18, 2020).

[Google Scholar](https://scholar.google.com/scholar?q=%E2%80%98WHO%20%7C%20Breast%20cancer%E2%80%99%2C%20WHO.%20http%3A%2F%2Fwww.who.int%2Fcancer%2Fprevention%2Fdiagnosis-screening%2Fbreast-cancer%2Fen%2F%20(accessed%20Feb.%2018%2C%202020).)

1. 11 Shen Q., Shao F., and Sun R., Prediction model of breast cancer based on xgboost, *Journal of Qingdao University (Natural Science Edition)*. (2019) **32**, no. 1.
2. Chen T. and Guestrin C., Xgboost: a scalable tree boosting system, *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, August 2016, San Francisco, CA, USA, 785–794.

[**Google Scholar**](https://onlinelibrary.wiley.com/action/getFTRLinkout?url=http%3A%2F%2Fscholar.google.com%2Fscholar%3Fhl%3Den%26q%3D%250AChen%2BT.%2Band%2B%250AGuestrin%2BC.%252C%2BXgboost%253A%2Ba%2Bscalable%2Btree%2Bboosting%2Bsystem%252C%2BProceedings%2Bof%2Bthe%2B22nd%2BACM%2BSigkdd%2BInternational%2BConference%2Bon%2BKnowledge%2BDiscovery%2Band%2BData%2BMining%252C%2BAugust%2B2016%252C%2BSan%2BFrancisco%252C%2BCA%252C%2BUSA%252C%2B785%25E2%2580%2593794.&doi=10.1155%2F2023%2F6530719&linkType=gs&linkLocation=Reference&linkSource=FULL_TEXT)

<https://dl.acm.org/doi/10.1145/2939672.2939785>

1. Breiman L., Random forests, *Machine Learning*. (2001) **45**, no. 1, 5–32, [**https://doi.org/10.1023/a:1010933404324**](https://doi.org/10.1023/a:1010933404324), 2-s2.0-0035478854

[**Google Scholar**](https://onlinelibrary.wiley.com/action/getFTRLinkout?url=http%3A%2F%2Fscholar.google.com%2Fscholar_lookup%3Fhl%3Den%26volume%3D45%26publication_year%3D2001%26pages%3D5-32%26journal%3DMachine%2BLearning%26author%3DL.%2BBreiman%26title%3DRandom%2Bforests&doi=10.1155%2F2023%2F6530719&doiOfLink=10.1023%2FA%3A1010933404324&linkType=gs&linkLocation=Reference&linkSource=FULL_TEXT)

1. Walker S. H. and Duncan D. B., Estimation of the probability of an event as a function of several independent variables, *Biometrika*. (1967) **54**, no. 1/2, 167–179, [**https://doi.org/10.2307/2333860**](https://doi.org/10.2307/2333860).
2. Cover T. and Hart P., Nearest neighbor pattern classification, *IEEE Transactions on Information Theory*. (1967) **13**, no. 1, 21–27, [**https://doi.org/10.1109/tit.1967.1053964**](https://doi.org/10.1109/tit.1967.1053964), 2-s2.0-84926662675.
3. Kaggle, <https://www.kaggle.com/datasets/kabshaansariya/breast-cancer>
4. Google,

<https://www.mdpi.com/2075-1729/14/8/1025>

1. Burnside ES, Liu J, Wu Y, Onitilo AA, McCarty CA, Page CD, et al. . *Acad Radiol.*2016;23(1):62–9. doi: 10.1016/j.acra.2015.09.007.

[[CrossRef](https://doi.org/10.1016%2Fj.acra.2015.09.007)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=+Acad+Radiol&title=+Comparing+Mammography+Abnormality+Features+to+Genetic+Variants+in+the+Prediction+of+Breast+Cancer+in+Women+Recommended+for+Breast+Biopsy&author=ES+Burnside&author=J+Liu&author=Y+Wu&author=AA+Onitilo&author=CA+McCarty&volume=23&issue=1&publication_year=+2016&pages=62-9&pmid=26514439&doi=10.1016/j.acra.2015.09.007&)]