Breast Cancer Classification and Prediction

Toshani Rungta
Computer Science Department
PES University
Bangalore, India
toshanirungta@gmail.com

Tankala Sunaina
Computer Science Department
PES University
Bangalore, India
tsunaina123@gmail.com

Susan Mathew K
Computer Science Department
PES University
Bangalore, India
susankakassery@gmail.com

Sanjana Murthy
Computer Science Department
PES University
Bangalore, India
sanjana2000murthy@gmail.com

***Abstract* — Breast cancer starts from the breast cells and progresses in stages. Early symptoms are often: new lump in the underarm or in breast, itching or discharge from the nipples and change in skin texture of the nipple or breast. Initially, the cancerous growth is "in situ" (i.e locally contained) where it generally causes no symptoms and has minimal potential for metastasis (spread). This situation is generally easily treatable. The fatality of the disease increases due to widespread metastasis which is a major cause of death of cancer patients. As of the end of 2020, there were 7.8 million women alive who were diagnosed with breast cancer in the past 5 years, making it the world's most prevalent cancer. These are the statistics despite having sufficient data to classify and predict the chances of this disease, thereby making it important for the medical field to start looking into the same. With the aim of accomplishing this, we have chosen a breast cancer dataset to verify the existing classification models. The chosen dataset is from Kaggle and has been pre-cleaned requiring minimum pre-processing. Only NaN values had to be dropped which has been achieved using pandas. It has 569 rows and 32 columns, with crucial information that helps in the classification process. We implemented various training models on it and verified the results to find that AdaBoost was the best classifier among the lot.**

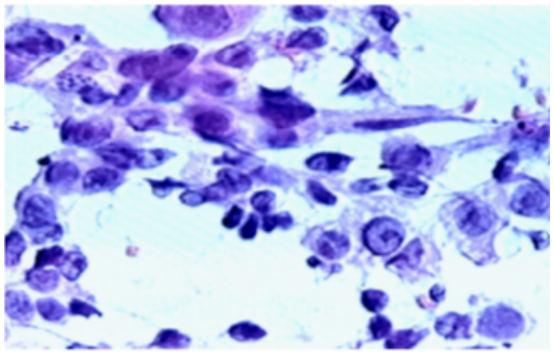

Figure 1: Breast cancer tumor cells.

I. INTRODUCTION

Cancer is one of the deadliest diseases in the world, it can develop almost anywhere in the body. There are over 100 different types of cancers that are currently known to us. Breast cancer is best treated if identified in its early stages. It, therefore, becomes a crucial task to differentiate the tumour cells from benign so that the patient can seek early medical help. It occurs when abnormal cells grow, divide, and create new cells that the body does not need and that do not function normally. These extra cells are what form a mass called a tumour.

Some tumours are 'benign' or non-carcinogenic. These tumours usually stay in one spot in the breast and do not cause any major health issues. Other tumours are 'malignant' or carcinogenic. As these tumours grow, they can spread throughout the breast or to other parts of the body.

Classification is the method of recognizing the group to which a new process belongs when the dependency is understood by the model using a training data set.

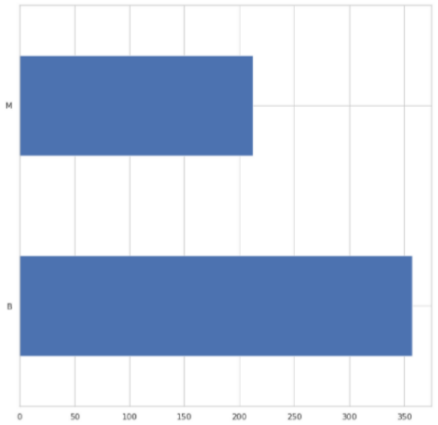

Figure 2: Diagnosis distribution.

II. PROBLEM STATEMENT

Breast cancer is a malignant tumour that occurs in females with the highest incidence, which has serious adverse effects on a woman's health. Therefore, early and accurate diagnosis of breast cancer patients is extremely crucial to recovery.
We chose the Wisconsin Breast Cancer(Diagnostic) Data Set for our study to classify and predict if the patient has breast cancer.

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430

Figure 3: Available attributes from the dataset

III. PREVIOUS WORK

There was research done for a prospective breast cancer risk prediction model for women undergoing screening mammography. Studies show that predictive models based on data from non-Hispanic White women showed the highest predictive accuracy. Research also shows that the models trained on data from different demographics performed well only when tested against the same demographic, this is important for us to know so we can select our dataset accordingly. New developments in this field have shown that some additional features of the tumour could be taken into account for a viable classification model, earlier work has not considered some of these features. Previous models have seen great success using SVM models for classification, but we found that using some ensemble methods have seen good results for similar problems.

IV. PROPOSED SOLUTION

A. PREPROCESSING

For our pre-processing process, we first identify the missing values and visualize the percentage of missing values from the dataset. Since our data consists of tumour attributes like radius and texture mean, perimeter, concavity and many more, we plot a graph to see how many tumours were classified as benign and how many are classified as malignant. A pie chart for the same is plotted for better visual representation. We next find the missing attributes in the dataset. We notice that one column contains 32 null values. We drop this column as it does not give us much insight. We did not have to do PCA as we could observe clear patterns in trends from our initial visualization and EDA. We went on to check for the correlation between the density of the recorded tumours vs their radius mean. A comparison between benign and malignant tumours was plotted. We observed that malignant tumours are less dense but have more radius mean than benign tumours. This gives a good initial metric to differentiate between the two types of tumours.

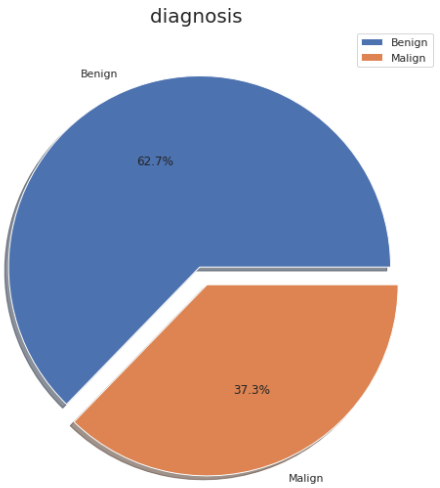

Figure 4: Pie Chart Distribution of Diagnosis

Next, we checked the correlation between radius mean and their respective diagnosis. It aligned with our initial metric from our previous plot. A pair plot helps to plot among the most useful features. We took attributes such as ‘diagnosis’, ‘radius_mean’, ‘texture_mean’, ‘perimeter_mean’, ‘area_mean’, ‘smoothness_mean’, ‘compactness_mean’, ‘concavity_mean’, ‘concave points_mean’, ‘symmetry_mean’, ‘fractual_dimension_mean’ to get further insights into the dataset.

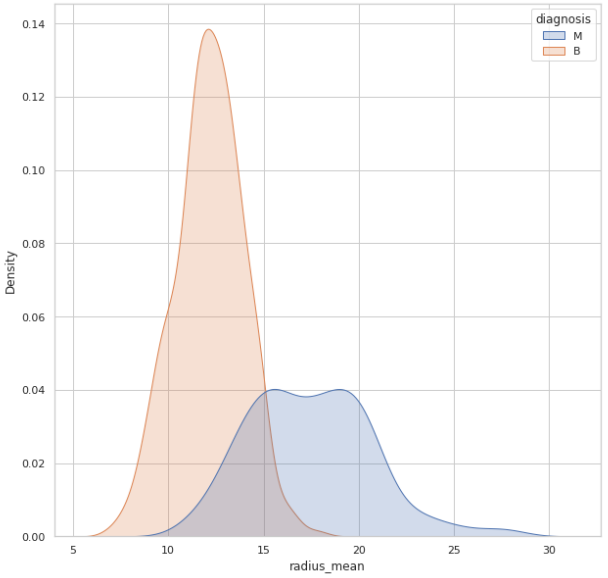

Figure 5: Density us radius_mean KDE plot.

We also plotted a correlation matrix to compare the relationship between each attribute to all other attributes. After extracting the usefulness of each attribute and its correlation to other attributes, we dropped the ‘less’ important features and plotted a correlation matrix again using the attributes left behind.

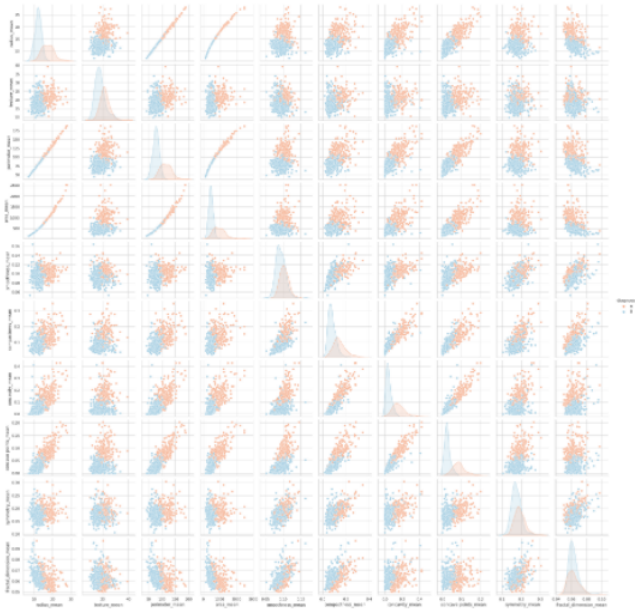

Figure 6: Pair plot of all the useful features.

B. TRAINING AND TESTING

We first split the data into training and testing data in the ratio of 70:30. This is done to train the models using the training dataset and then check the model's accuracy using the testing dataset.

The first model that we implemented was logistic regression. This model is used when a binary dependent variable is present. Here, logistic regression model is used to discriminate between benign and malignant breast tumors and identify important features associated with the same.
We got an accuracy of 58.82%. Since this is only slightly better than random guessing. Therefore, we explored other classification models.

Bayes naive algorithm is a supervised learning algorithm, based on Bayes' theorem. It is used to solve classification problems. It assumes that the occurrence of a certain characteristic is independent of the occurrence of another. It is a probabilistic classifier, which means that the prediction is based on the probability of the object. Naïve Bayes classifier gave a training score of 63.2967% with an F1 score of 0.74 and a recall of 0.99

Next, we explored kNeighborsClassifier. This is a supervised machine learning algorithm that can be used to solve both classification and regression problems.KNN is one of the simplest algorithms used in Machine Learning for regression and classification problems. KNN algorithms use data and classify new data points based on similarity measures. Classification is done by a majority vote to its neighbors.This model gives us the best output when the number of nearest neighbors considered is 7. We obtained a training score of 70.18% which was only slightly better than the Naïve Bayes classifier.

Next, we explored the decision tree classifier where we set the max depth as 6 and the random rate to '123'. It is a form of supervised machine learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter and this continues till we reach a conclusion. With these parameters, we got an accuracy of 94.7368% which is much better than any of the previous models we have trained so far. The F1 score was 0.95 and recall was 0.94.

Next, we decided to see if other classification models back up this accuracy or if the model was overfitting. So we did a Random Forest classification which gave an accuracy of 97.368%. It is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression and is based on the concept of ensemble learning - a process of combining multiple weak classifiers to solve a complex problem, with an improved performance of the model. Random Forest is a classifier which contains a certain number of decision trees over various subsets of the given set and takes the average to improve the precision of that set from data. Instead of relying on a decision tree, the random forest takes the size of each tree and predicts the final output based on the majority of the predictions. The F1 score was 0.98 and recall was 0.97, which is close enough to the decision tree classifier.

Lastly, we wanted to try an ensemble method learner, so we implemented the AdaBoost classifier. AdaBoost, also known as Adaptive Boosting, is a Ensemble learning technique. The most commonly used algorithm with AdaBoost is one-level decision trees, that is, 1-division decision trees. These trees are also called Decision Stumps. It builds a model and gives all data points equal weights .It then assigns higher weights to points that are wrongly classified. Now all the points that have more weights are given more importance in the next model. It will keep training models until a low error is received. This model gave us an accuracy of 98.2456% with an F1 score of 0.99 and a recall rate of 1.00

V. CHALLENGES

1) Breast cancer is a highly heterogeneous disease. Thus, one of the major challenges is building accurate and computationally efficient algorithms for classifying patients to guide therapeutic decision making at the point of care.
2) A major challenge with breast cancer is identifying patients at risk of recurrence. Once a patient has relapsed, their chances of survival decrease significantly. For example, 73% of patients who develop metastasis will die after 5 years.
3) Considering the fatality of the disease, wrong predictions could lead to potential complications and even death of the patient and hence accuracy of prediction was a major concern.

VI. CONCLUSION

As mentioned in the previous section, we have implemented Logistic regression, Decision Tree classifier, Random Forest classifier, KNeighbors classifier, Adaboost classifier, Gradient Boosting classifier and Naive Bayes models. On computing the confusion matrix and accuracy for each of the models, we found that the **Adaboost Classifier** model gave the highest accuracy of **98.24%** as it is an ensemble learning model where it uses several weak learners but ensures that the final model has a good performance metric without overfitting or underfitting. The **Random Forest** model is a close second with an accuracy of **96.57%**.

VII. ACKNOWLEDGEMENT

We thank the CSE Department, PES University for giving us the wonderful opportunity to do a project on machine learning. We further thank and are grateful for the guidance and support of our mentor Prof. R. Bharathi. Last but not the least, we thank our

fellow students who always gave us immaculate and important suggestions at each step. This has been an insightful educational experience and extends our heartfelt gratitude for the same.

REFERENCES

[1] https://ieeexplore.ieee.org/document/9421338
[2] https://ieeexplore.ieee.org/document/6016771
[3] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0226765
[4] https://ieeexplore.ieee.org/document/8965528
[5] https://ieeexplore.ieee.org/abstract/document/8820378
[6] https://ieeexplore.ieee.org/document/5703994
[7] http://ijiepr.iust.ac.ir/article-1-1069-fa.pdf
[8] https://ieeexplore.ieee.org/document/8605180
[9] https://pubs.rsna.org/doi/full/10.1148/radiol.2019182716
[10] https://academic.oup.com/jnci/article/98/17/1204/2521747?login=true
[11] https://core.ac.uk/download/pdf/295538238.pdf
[12] https://www.jigsawacademy.com/blogs/data-science/classification-and-prediction-in-data-mining/
[13] https://www.who.int/news-room/fact-sheets/detail/breast-cancer
[14] https://owkin.com/cancer/owkin-breast-cancer-treatment-response-recurrence/#:~:text=On%20the%20other%20hand%2C%20the%20challenge%20we%20face,who%20develop%20metastasis%20will%20die%20after%205%20years.
[15] https://medium.com/swlh/predicting-breast-cancer-using-logistic-regression-3cbb796ab931