Breast Cancer Detection

Smruthi Gowtham
Computer Science
PES University
Bangalore, India
smruthig01@gmail.com

Snigdha Sinha
Computer Science
PES University
Bangalore, India
snigdhasinha0811@gmail.com

Vridhi Goyal
Computer Science
PES University
Bangalore, India
goyalvridhi@gmail.com

Yashi Chawla
Computer Science
PES University
Bangalore, India
yashichawla1@gmail.com

***Abstract*—Breast cancer is a common incursive cancer among women and one of the main causes of cancer death, which can be classified as Benign or Malignant. Diagnosis of breast cancer is time-consuming and due to its importance, it is imperative to develop a system that can automatically diagnose breast cancer in its early stages so that it can be treated efficiently. Breast Cancer prediction aims at extracting features from the given samples and predicting it as Benign or Malignant. The dataset chosen is extracted from the Wisconsin Breast Cancer Dataset. This implementation uses many different approaches and compares the accuracy of the various models through performance metrics.**

Keywords— cancer, prediction, learning, data mining, support vector machines, artificial neural network, k nearest neighbors

I. Introduction

Breast Cancer is one of the most prevalent and deadliest cancers that exist, especially for women. It comes with a copious amount of both physical and psychological damage. Breast cancer arises in the lining cells of the ducts in the glandular tissue of the breast. Initially, the cancerous growth is confined to the duct where it generally causes no symptoms and has minimal potential for spread.

Over time, these in stage 0 cancers may progress and invade the surrounding breast tissue then spread to the nearby lymph nodes or other organs in the body. If a woman dies from breast cancer, it is because of widespread metastasis. There are many risk factors, when it comes to breast cancer such as race, age, genes, exercise level, alcohol consumption, etc. and there are different types of breast cancer with different stages, spread, and aggressiveness.

Breast Cancer is the most common type of medical hazard found in middle-aged women. In 2020, over 2.3 million women were diagnosed with breast cancer with a death rate of 29.78 percent. As of 2020, there are 7.8 million women alive who were diagnosed with breast cancer in the past 5 years. There are more lost disability-adjusted life years (DALYs) [8] by women to breast cancer than any other type of cancer.

Breast cancer treatment includes surgical removal, radiation therapy, etc, and can be highly effective, if it can be detected at an early stage. Misdiagnoses can lead to improper treatments where people lose their prime time for treatment. And that's why Breast cancer prediction has long been regarded as an important research problem in the medical community. It would be very useful to have a system that would allow for early detection and prevention which would increase the survival rates of people with breast cancer.

The Breast Cancer Wisconsin Dataset was obtained from the UCI Machine learning repository and is a benchmark dataset. It contains 569 instances and 32 features that provide precise information pertaining to the occurrence of breast cancer. They are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass and describe the characteristics of the cell nuclei present in the image. The dataset is distributed over 37.25% cancerous samples and 62.75% non-cancerous samples.

Table I: Summary of Wisconsin Breast Cancer

Attribute Number	Attribute Description	Range		
		Mean	Standard error	Largest value
1	Radius	6.98 - 28.11	0.11- 2.87	7.93 - 36.04
2	Texture	9.71 - 39.28	0.36 - 4.89	12.02 - 49.54
3	Perimeter	43.79 - 188.50	0.76 - 21.98	50.41 - 251.20
4	Area	143.50 - 2501.00	6.80 - 542.20	185.20 - 4254.00
5	Smoothness	0.05 - 0.16	0.00 - 0.03	0.07 - 0.22
6	Compactness	0.02 - 0.35	0.00 - 0.14	0.03 - 1.06
7	Concavity	0.00 - 0.43	0.00 - 0.40	0.00 - 1.25
8	Concave points	0.00 - 0.20	0.00 - 0.05	0.00 - 0.29
9	Symmetry	0.11 - 0.30	0.01 - 0.08	0.16 - 0.66
10	Fractal dimension	0.05 - 0.10	0.00 - 0.03	0.06 - 0.21

The dataset was scrutinized for unknown values, inconsistency, and erroneous data as they can have a consequential effect on the interpretations that can be derived from the data. No unknown instances were discovered. Outliers were found but were not removed or replaced as they can give valuable insights.

With the help of the latest, efficient and advanced screening methods, the majority of such cancers are diagnosed when the disease is still at a localized stage. The

utility of machine learning techniques in healthcare analysis is growing progressively. Certainly analysis of a patient's clinical data and physician's judgment is the most considerable feature in diagnosis. Most of the possible medical flaws can be avoided by using classification systems and also offer healthcare data to be analyzed in less time and in a more exhaustive manner. Accurate and timely prediction of breast cancer allows physicians and healthcare providers to make the most favorable decision about the patient's treatment.

II. Literature review

Data mining is the task of processing data and discovering hidden information. The urgent need to transform data into knowledge and information has gained greater attention over the last few years. The two widely used approaches in data mining are classification and clustering. The purpose of the study is to design an accurate machine learning model that predicts breast cancer at early stages. Our study focuses on the prediction of breast cancer, from the area of clinical medicine. This section consists of a review of work done in supervised learning algorithms and their applications to solve the given problem based on different approaches. Amongst the largely available reports on medical data analysis utilizing machine learning and data mining, all with good accuracies, the below-mentioned reports are just representative.

A. GRU-SVM

Abien Fred M. Agarap's study introduces the recently-proposed GRU-SVM [1], which is a gated recurrent neural network along with an SVM to be considered to solve the given problem. The study includes other approaches such as linear regression, multilayer perceptron, nearest neighbor, softmax regression, and support vector machine. The Wisconsin Breast Cancer Database is linearly separable. However, the GRU-SVM also gives commendable results with a mid-level performance with an accuracy of 93.75%, which can be explained due to the sensitivity of RNNs to weight initialization [2], as they are arbitrarily assigned and due to the non-linearities [3] which were introduced due to the mechanisms of gating. However, the linear classifiers perform better as the utilized dataset in hand is linearly separable. To find optimal hyper-parameters for the ML algorithms, a CV technique such as k-fold cross-validation was employed.

B. Random Forest

The study by Yixuan Li, Zixuan Chen [4], compares five classification models. The study explores the relationship between breast cancer and some attributes so that the death probability of breast cancer can be reduced. The study compares Decision Tree, Random Forest, Support Vector Machine, Neural Network, and Logistics Regression, on two different datasets Breast Cancer Coimbra Dataset (BCCD) and Wisconsin Breast Cancer Database (WBCD). And it concludes Random forest is selected compared to the other four models, attaining accuracy of 96.1%, and F-measure metric of 95.5% followed by Decision Trees, with an accuracy of 96.1% and F-measure metric of 94.1% The study further suggests that in order to gain more efficiency and accuracy, the RF model can be combined with other data mining technologies. The study performs prediction 50 times on different randomly split training and test data splits for the purpose of verification. The study also showcases some limitations. And suggests that the lack of raw data affects the accuracy.

C. Multilayer Perceptron

The Paper by Mohammed Abdul Hay Abu Bakr, Haitham Maher Al-Attar, Nader Kamal Mahra, Samy S. Abu-Naser [5], uses the prediction power of a neural network to classify whether Breast Cancer is a benign or malignant cancer. They ran a series of tests to determine the number of hidden layers and neurons in each hidden layer. They used the JustNN environment for building the network that was a feed-forward Multilayer Perceptron with one input layer, one hidden layer with 10 neurons, and one output layer. The average predictability rate achieved was 99.57%.

D. PCA

The study by Haifeng Wang and Sang Won Yoon [6] discusses breast cancer feature extraction based on data mining methods. It compares 4 data mining models, support vector machine (SVM), artificial neural network (ANN), Naive Bayes classifier, AdaBoost tree as well as eight hybrid models that are tested on two data sets, Wisconsin Breast Cancer Database (1991) and Wisconsin Diagnostic Breast Cancer (1995). PCA, as a dimension reduction technique, manifests some advantages in terms of prediction accuracy and efficiency. A combination of scree-plots is applied in this research to decide the necessary amount of PC. The results are used to build eight new hybrid models PCs-SVM, PCi-SVM, PCs-ANN, PCi-ANN, PCs-Naïve, PCi-Naïve, PCsAdaBoost, and PCi-AdaBoost, in which PCs represent selected principal components based on scree plot criteria, and PCi denotes the principal components selected based on 95% of correlation explained, to test whether the dimension reduction method can influence prediction effectiveness and efficiency or not. Then a series of paired *t*-test were performed to compare the prediction accuracy for each model and the one with the highest accuracy was finalized, which turned out to be PCi-ANN.

E. Artificial Neural Networks

This paper by Milon Islam, Md. Rezwanul Haque, Hasib Iqbal, Md. Munirul Hasan, Mahmudal Hasan and Md. Nomani Kabir [7] uses the Wisconsin Breast Cancer Dataset and compares five supervised machine learning algorithms of Support Vector Machine (SVM), Logistic Regression K-Nearest Neighbors, Artificial Neural Networks (ANNs),

and Random Forests. They have used performance metrics such as accuracy, sensitivity, specificity, negative predictive value, false negative rate, false positive rate, F1 score, and Matthews Correlation Coefficient to measure the performance of their models. Furthermore, these algorithms were assessed on the precision-recall area under curve and ROC curve. From these metrics, they concluded that ANNs provided the highest accuracy, precision, and F1 score of 98.57%, 97.82%, and 0.9890, respectively, followed by SVM which gave 97.14%, 95.65%, and 0.9777 accuracy, precision, and F1 score respectively.

III. Problem Statement

Our study focuses on the prediction of breast cancer (malignant or benign) by extracting features from the samples obtained through the Wisconsin Breast Cancer Database (WBCD). The solution aims to assess whether a lump in a breast could be malignant (cancerous) or benign (non-cancerous) by examining the radius, perimeter, concavity, etc. features of the sample. Each sample contains the input features and target variables of a different patient. The data set is divided into training and testing data. 80% of the instances will be assigned for training, and 20% for testing. Malignant tumors represent 38% of the samples, and benign tumors represent 62% of the samples approximately.

IV. Our Approach

Through the literature survey, we have analyzed the work done in this field by other scholars. We observe that machine learning algorithms such as Support Vector Machines (SVM), K- Nearest Neighbours (KNN), and Neural Networks (NNs) are among the most effective methods for the prediction of breast cancer. Therefore, we plan to train and test our dataset for breast cancer prediction by building supervised ML models on SVM, KNN, and NN. We will test their performance measures through performance metrics such as accuracy, precision, recall, F1 score, and other methods which seem necessary over time.

V. Exploratory Data Analysis and Visualisation

We performed certain checks on the dataset before proceeding with the Visualisation, such as checking for null values, duplicates, and inconsistent data.

No unknown values were discovered in the data.

```
df.isnull().sum()
```

```
id                         0
diagnosis                  0
radius_mean                0
texture_mean               0
perimeter_mean             0
                          ..
compactness_worst          0
concavity_worst            0
concave points_worst       0
symmetry_worst             0
fractal_dimension_worst    0
Length: 32, dtype: int64
```

No duplicate values were detected either.

```
df["id"].is_unique
```

```
True
```

The target variable in the dataset only contains the values M and B, which stands for Malignant and Benign respectively. Hence, there are no inconsistencies in the data.

```
df['diagnosis'].unique()
```

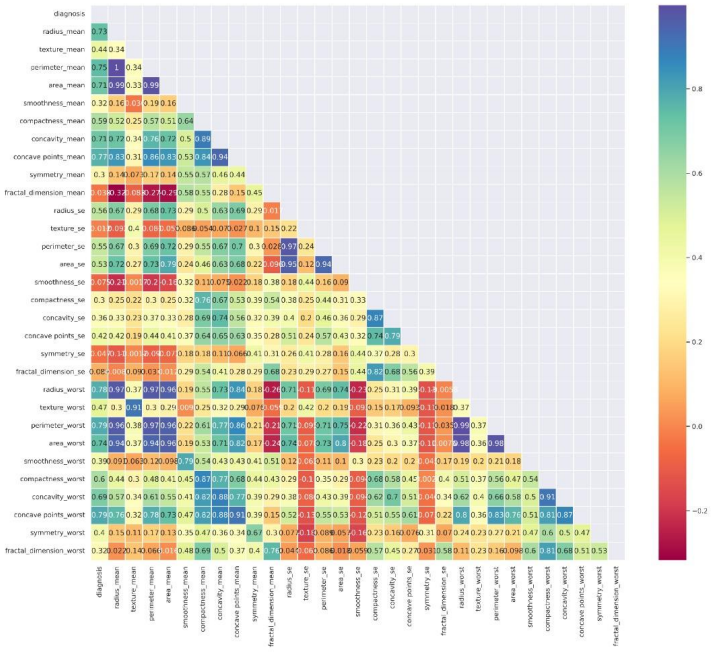
```
array(['M', 'B'], dtype=object)
```

No negative values were detected in the dataset.

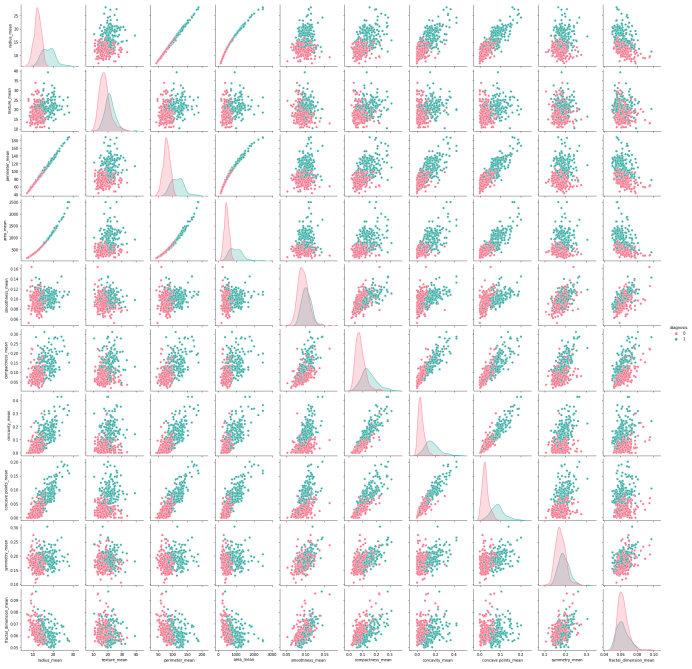
```
(df.iloc[:,2:] < 0).any().any()
```

```
False
```

We ran a series of inputs-target correlations that might indicate to us what factors most influence a tumor to be malignant or benign. It was observed that attributes Radius and Concavity showed a high positive correlation to the target variable diagnosis. Perimeter_mean and radius_mean had a correlation coefficient of one which indicates a high linear correlation, so we decided to drop perimeter_mean. Similarly, between Area_mean and Radius_mean, Area_mean was dropped.



We plotted a pair plot to analyze the pairwise relationships of the features in the dataset with target value diagnosis as the hue. It was observed that almost all the values in the plot are linearly separable against their respective target values.



The dataset was normalized to ensure that it looks and reads the same way across all records. We applied PCA to the data to get a better understanding of the visualization. Based on elementary analysis of the logical meaning of each of the attributes, we can reasonably claim that the columns that give the "mean values" hold more information for our analysis compared to the "standard errors" and "worst" values of each. However, we will be retaining these attributes till we gain a better understanding of the domain.
PCA converts feature space into uncorrelated variables based on a linear function. In terms of nonlinear feature reduction methods, we plan to test some other techniques such as *k*-means. Furthermore, all the papers we referred to [6], used standard datasets, but for a more refined analysis, we plan to study some raw datasets such as SEER.

References

[1] Abien Fred Agarap. 2017. A Neural Network Architecture Combining Gated Recurrent Unit (GRU) and Support Vector Machine (SVM) for Intrusion Detection in Network Traffic Data. arXiv preprint arXiv:1709.03082 (2017).

[2] Abdulrahman Alalshekmubarak and Leslie S Smith. 2013. A novel approach combining recurrent neural networks and support vector machines for time series classification. In Innovations in Information Technology (IIT), 2013 9th International Conference on. IEEE, 42–47

[3] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014).

[4] Yixuan Li, Zixuan Chen. Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction. Applied and Computational Mathematics. Vol. 7, No. 4, 2018, pp. 212-216. doi: 10.11648/j.acm.20180704.15

[5] Bakr, Mohammed Abdul Hay Abu; Al-Attar, Haitham Maher; Mahra, Nader Kamal & Abu-Naser, Samy S. (2020). Breast Cancer Prediction using JNN. *International Journal of Academic Information Systems Research (IJAISR)* 4 (10):1-8.

[6] Wang, Haifeng & Yoon, Sang Won. (2015). Breast Cancer Prediction Using Data Mining Method.

[7] Islam, M.M., Haque, M.R., Iqbal, H. *et al.* Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. *SN COMPUT. SCI.* 1, 290 (2020).

[8] https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[9] https://www.neuraldesigner.com/learning/examples/breast-cancer-diagnosis