Breast Cancer Detection

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

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

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

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

***Abstract*——Breast cancer is a prevalent incursive cancer among women. It is one among the primary causes of death related to cancer, which can be classified as Malignant or Benign. Breast cancer diagnosis is time consuming and due to its gravity, it is imperative to design a solution to automate the process of identification of the same 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 compares six basic models namely, Logistic Regression, SVC, AdaBoost, Neural Network Nearest Neighbours and Random Forest through performance metrics like, F1-score and F1-stratified score. The results reveal that the highest performing model for this problem statement is Logistic Regression with an F1-score of 0.9788 and F1-stratified score of 0.9822.**

***Index Terms*—Breast Cancer, Prediction, Regression, AdaBoost, neural network, nearest neighbors**

I. Introduction

Breast Cancer originates in the lining cells of the glandular tissue ducts of the breast. Initially, the cancerous growth is limited to the duct where it usually does not show any symptoms and has very little scope for spreading. Gradually, the cancerous cells in stage 0 cancers may advance and invade the neighbouring breast tissue and then progress to the adjacent lymph nodes or other organs in the body within reach. Death due to Breast Cancer 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 multiple types of breast cancer with differing spread, stages, and aggressiveness.

Breast cancer treatment includes surgical removal, radiation therapy, etc, and can be highly effective, if detected earlier and treated accordingly. Misdiagnoses can lead to improper treatments where people lose their prime time for treatment which is why Breast Cancer prediction has been a principal topic for research in the medical field. Implementation of a prediction model which would help in early detection of breast cancer will be very helpful in increasing the survival rates of the patients diagnosed.

The dataset that has been used in our project is The Wisconsin Breast Cancer Dataset (WBCD), acquired from the repository of UCI Machine learning, is a benchmark dataset. The dataset is distributed over 37.25% cancerous samples and 62.75% non-cancerous samples.

TABLE I
Summary of Wisconsin Breast Cancer Dataset

Attribute Description	Range		
	Mean	***Standard Error***	***Maximum Value***
Radius	6.98 - 28.11	0.11 - 2.87	7.93 - 36.04
Texture	9.71 - 39.28	0.36 - 4.89	12.02 - 49.54
Area	143.50 - 2501.00	6.80 - 542.20	185.20 - 4254.00
Perimeter	43.79 - 188.50	0.76 - 21.98	50.41 - 251.20
Smoothness	0.05 - 0.16	0.00 - 0.03	0.07 - 0.22
Concavity	0.00 - 0.43	0.00 - 0.40	0.00 - 1.25
Concave Points	0.00 - 0.20	0.00 - 0.05	0.00 - 0.29
Compactness	0.02 - 0.35	0.00 - 0.14	0.03 - 1.06
Fractal Dimension	0.05 - 0.10	0.00 - 0.03	0.06 - 0.21
Symmetry	0.11 - 0.30	0.01 - 0.08	0.16 - 0.66

Statistics report that in the year 2020, over 2.3 million females were found to be diagnosed with breast cancer with a death rate of 29.78 percent. According to a 2020 survey, breast cancer was detected in 7.8 million women in the past 5 years. There are more Disability-Adjusted Life Years (DALYs) [1] lost by women to breast cancer than any other form of cancer.

In this Analysis, we aim to come up with an improvised approach which helps in effective and accurate prediction of malignant and benign breast cancer.

II. Previous Work

The paper titled 'On Breast Cancer Detection: An Application of Machine Learning Algorithms on the

Wisconsin Diagnostic Dataset' by Abien Fred M. Agarap[2] compares Machine Learning algorithms of Linear Regression, Support Vector Machine (SVM), Softmax Regression, GRU-SVM, Nearest Neighbour search and Multilayer perceptron (MLP). Out of the six models mentioned, MLP emerges as the highest performing model with an accuracy of approximately 99.04%. However, the GRU-SVM also gives commendable results with an accuracy of 93.75%, which can be explained due to the sensitivity of RNNs to weight initialization[3], as they are arbitrarily assigned and due to the non-linearities[4] which were introduced due to the mechanisms of gating.

In the paper titled 'Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction'[5], by Zixuan Chen and Yixuan Li, five ML models, namely Random Forest, Decision Tree , Neural Network, Logistic Regression and SVM have been compared through indicators like accuracy values, AUC values and F-measure metric. The comparison showed that Random Forest performs better than the other four models and yields the highest accuracy score of 96.1% and F-measure metric of 94.1%.

While the two papers succeed in achieving high model accuracy, they do not talk about cross validation techniques which can provide a more accurate model prediction.

However, this approach has been discussed in the study by Sang Won Yoon and Haifeng Wang in their paper 'Breast Cancer Prediction Using Data Mining Method.'[6] It discusses breast cancer feature extraction based on data mining methods and presents a comparison among four models namely, Naive Bayes Classifier, AdaBoost Tree, Artificial Neural Network (ANN) and SVM. To approxiamte the error of these models, 10-fold cross-validation method is implemented. It also compares eight hybrid models namely, PCs-SVM, PCi-SVM, PCs-ANN, PCi-ANN, PCs-Naïve, PCi-Naïve, PCsAdaBoost, and PCi-AdaBoost. A sequence of paired t-tests are performed to compare the prediction accuracy for each model and PCi-ANN turned out to be the one with the highest accuracy.

The cross validation technique has also been implemented by Md. Milon Islam et. al in the paper titled 'Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques'[7]. They have used a ten-fold cross-validation technique on the Wisconsin Breast Cancer Dataset (nine-fold for training and one for testing the model) and have performed a comparison between five ML algorithms, namely, SVM, Random Forest, ANN, Logistic Regression and K Nearest Neighbours. The performance of the models were measured using metrics like specificity, accuracy, Matthews Correlation Coefficient and F1 score. Judging by these metrics, ANN was found to be the best model with an accuracy of 98.57%, a precision of 97.82% and an F1 score of 98.9%.

III. Proposed Solution

This study's primary aim is binary classification of the given samples of cell nuclei as benign or malignant. For the analysis of our problem statement, six ML models have been used which are, Logistic Regression, Support Vector Classifier (SVC), AdaBoost, Neural Network, K-Neighbours classifier and Random Forest.

A. Logit Model

Logit Model, also called Logistic Regression is used for the purpose of binary classification via probability of belonging to either of the classes.It is used for finding existence of a relationship between a discrete dependent variable and several independent variables. It can be generalized to multi-class classification. Rather than Mean Squared Error (MSE), Logistic Regression uses Maximum Likelihood Estimate (MLE) as the loss function.

B. Support Vector Classifier

Support Vector Machine (SVM) is a popular machine learning algorithm that falls under supervised ML category. It can be used for classification problems as well as regression problems. In SVM, each feature for a particular data point represents one coordinate. If f is the number of attributes, the combination of all features results in the data point being plotted as a single point in f-dimensional space. For the purpose of classification, a hyperplane is chosen such that it maximises the margin of difference between the classes.

C. Adaptive Boosting

Adaptive Boosting, often abbreviated as AdaBoost, is an ensemble learning method. It is one of the earliest boosting algorithms. It combines several weak learners into one single Model. The weak models used in this technique are classifiers with a little over 50 percent accuracy. The working methodology is: it assigns a higher weight to misclassified data points and reduces the weight for the correctly classified ones.This ensures that the subsequent classifiers give higher priority to the instances that were previously misclassified.

D. Neural Network

A Neural Network is a bio-inspired model of Machine Learning that is formed by connecting units known as artificial neurons. It comprises of layer of nodes which can be classified into three different containers: An input layer, hidden layers and an output layer. Nodes are interconnected and are associated with weights and biases. When the output of a node, post activation function, crosses the threshold value, the node gets activated. Such a node then forwards the data to the next layer. A node which has not been activated will not pass the data on. It can infer a model even for data that is not linearly separable.

E. K Nearest Neighbours

K Nearest Neighbours (KNN) is a supervised learning algorithm. It is an instance based algorithm as it doesn't build a model based purely on the training data and waits for the test instance to do the same. For the given problem statement, KNN has been used for classification purpose. The inductive bias for KNN is that, given a new data point with an unknown class, it tends to belong to the same class as the bulk in its immediate neighbourhood.

F. Random Forest

Random Forest is an ensemble learning technique which combines several models to come up with solutions to problems such as classification. The algorithm comprises of several decision trees. The 'forest' generated is trained through bagging or bootstrap aggregation. It produces a reasonable prediction without hyperparameter tuning. It provides an outcome which relies on the prediction of the decision tree while eradicating all it's limitations.

IV. Workflow

The Wisconsin Breast cancer Dataset contains 569 instances and 32 features that provide accurate information regarding the diagnosis of breast cancer. The columns of the dataset represent the features of the cell nuclei found in the digitised image of fine needle aspirates (FNA) of breast mass.

A. Data Cleaning

During the process of data cleaning, the dataset (WBCD) was tested for inconsistencies, duplicates and null values and was found to be a clean dataset.

B. Data Preprocessing

For the types of models that have been trained for the given problem statement, it has been found that building models from the raw data without preprocessing yields poor results. Hence, multiple combinations of standardization, normalization and Principal Component Analysis (PCA) have been implemented as a part of data preprocessing. Furthermore, it was found that unstandardized data also yields relatively poor results. Hence the various combinations used were as follows:

TABLE II
Combination of Preprocessing Techniques

Standardize	***Normalize***	***PCA***
True	True	15 components
True	True	False
True	False	15 components
True	False	False

C. Attribute Dependencies and Correlation

To reduce the dimensionality of our dataset, the correlation between the features was studied and the output was observed. A series of input-target correlations were run that were indicative of 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.

D. Evaluation Metrics and Cross Validation

For our analysis, different evaluation metrics such as accuracy, F1 score and F1-stratified score have been proposed. For the purpose of testing the performance of the classifiers on varied sets of data, K-Fold cross-validation technique has been adopted across the entire dataset. This helps to verify that introduction of high bias does not make the model prone to underfitting. For the purpose of ensuring equal distribution of malignant and benign cases, both stratified and non stratified splitting has been used.

E. Classification Models

For the purpose of observing how different classifiers solve this issue, classifiers working on different approaches were proposed - Logistic Regression, Support Vector Classifier (SVC), AdaBoost, Neural Network and K-Neighbours classifier. GridSearch was performed with different hyperparameter values on all the classifiers used, in order to obtain a model with optimal performance.

1) Logistic Regression: Logistic Regression was chosen as the first classifier model. A cross validated Grid Search was used for the purpose of tuning the model. This enabled it to try out various combinations of parameters and allowed it to choose the optimal model.

TABLE III
Possible Parameters for Grid Search

Attributes	***Possible Values***
Penalty	l1, l2, elasticnet
C	0.5, 0.4, 0.3, 0.2, 0.1, 1, 10
Tol	1e-1, 1e-2, 1e-3, 1e-4, 1
Solver	"liblinear", "lbfgs"
Dual	True, False

TABLE IV
Chosen parameters after Grid Search

Attributes	***Chosen Values***
Penalty	l2
C	0.4
Tol	1
Solver	"liblinear"
Dual	True

2) SVC: The next chosen classifier was SVC. GridSearch was again implemented to choose the best set of hyperparameters which yields the optimum results.

TABLE V
POSSIBLE PARAMETERS FOR GRID SEARCH

Attributes	*Possible Values*
kernel	”linear”, ”rbf”, ”poly”, ”sigmoid”
C	0.5, 0.4, 0.3, 0.2, 0.1, 1, 10
degree	2, 3, 4, 5
gamma	scaled, auto
shrinking	True, False
tol	1e-1, 1e-2, 1e-3, 1e-4, 1
coef0	0.5, 0.4, 0.3, 0.2, 0.0, 0.1, 1, 10

TABLE VI
CHOSEN PARAMETERS AFTER GRID SEARCH

Attributes	*Chosen Values*
kernel	”poly”
C	0.1
degree	3
gamma	auto
shrinking	True
tol	1e-1
coef0	10

3) Adaboost: The third chosen classifier was AdaBoost. The best hyperparameters were chosen by performing a grid search on the possible values for each parameter while varying the depth of the decision tree.

TABLE VII
POSSIBLE PARAMETERS FOR GRID SEARCH

Attributes	*Possible Values*
$n_estimators$	10, 50, 100, 250, 500, 1000
$learning_rate$	0.1, 0.3, 0.5, 1, 5, 10
algorithm	”SAMME”, ”SAMME.R”
$base_estimator$	None, RandomForestClassifier

TABLE VIII
CHOSEN PARAMETERS AFTER GRID SEARCH

Attributes	*Chosen Values*
kernel	”poly”
C	0.1
degree	3
gamma	auto
shrinking	True
tol	1e-1
coef0	10

4) Neural Network: Neural Network is the fourth classifier that was used. To tune the model, the number of layers and the neurons per layer were modified continuously till the best result was obtained. ReLu was chosen as the activation function for the input layer as well as each hidden layer. The activation function chosen for the output layer was sigmoid. Adam was chosen as the optimizer. As the classification problem is a binary one, the loss function chosen was binary cross entropy.

TABLE IX
CHOSEN PARAMETERS AFTER GRID SEARCH

Attributes	*Chosen Values*
Layer Dimensions	128, 64, 1
Optimizer	Adam
Loss	binary cross entropy
Metric	accuracy
Epochs	200

5) KNN: The next chosen classifier model was K Nearest Neighbours. The model was tuned by varying the value of K and the chosen distance metric to determine the best result. To improve the model results, weights were also introduced. The grid search algorithm was used to implement the same.

TABLE X
POSSIBLE PARAMETERS FOR GRID SEARCH

Attributes	*Possible Values*
n_neighbors	2, 5, 10, 15, 20, 50
weights	distance, uniform
algorithm	brute, auto, kd tree, ball tree
leaf_size	1, 2, 4, 6, 8, 10, 20, 30, 40, 50
p	1, 2, 3
metric	euclidean, manhattan, minkowski, mahalanobis, cosine

TABLE XI
CHOSEN PARAMETERS AFTER GRID SEARCH

Attributes	*Chosen Values*
weights	uniform
n neighbors	5
leaf size	1
p	1
metric	manhattan
algorithm	auto

6) Random Forest: For the final model, the Random Forest classifier was trained. The Random Forest, being a powerful model in itself, does not require a lot of fine tuning to achieve high accuracy. Each decision tree in the forest was tuned based on it’s parameters. Grid search was deployed in order to choose the optimal set of hyperparameters which yielded the best outcome.

TABLE XII
POSSIBLE PARAMETERS FOR GRID SEARCH

Attributes	*Possible Values*
criterion	entropy, gini
max features	None, auto, log2
bootstrap	False, True
min samples leaf	10, 8, 6, 4, 2, 1
min samples split	10, 8, 6, 4, 2, 1
n_estimators	10, 50, 100, 250, 500, 1000

V. EXPERIMENTAL RESULTS

The six classifiers trained for the purpose of solving the aforementioned problem were Logistic Regression, SVC, AdaBoost, Neural Network, KNN and Random Forest. For each

TABLE XIII
CHOSEN PARAMETERS AFTER GRID SEARCH

Attributes	*Chosen Values*
bootstrap	False
min samples leaf	2
min samples split	10
n estimators	250
criterion	entropy
max features	auto

of these models, four possible pipelines of preprocessing techniques have been used, as mentioned in section IV B. The metrics used for evaluating each of these models for all the different pre-processing pipelines are F1 score and stratified F1 score. The accuracy, roc auc score and Mean Squared Error have also been calculated. The results obtained for the best model of each classifier, is as follows.

TABLE XIV
PERFORMANCE METRICS OF THE MODELS USED

Models Used	**F1 Score**	**MSE**	**Accuracy**	**ROC**
SVC	0.987342	0.093659	0.991228	0.987500
Logistic Regression	0.974359	0.132453	0.982456	0.980513
Random Forest	0.962025	0.162221	0.973684	0.968243
Adaboost	0.950000	0.187317	0.964912	0.956565
KNN	0.961039	0.162221	0.973684	0.973684

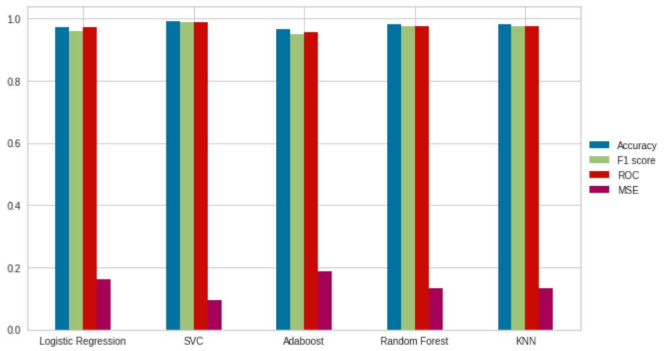

Fig. 1. Performance metrics of the models used

The best result was yielded by the Logistic Regression Classifier on a pre-processing pipeline consisting of standardization and PCA reducing the feature set to 15 components. The F1 and F1-stratified scores obtained for the above model were 0.9822 and 0.9788 respectively.

CONCLUSION

Analysis of the Data gave us a deeper understanding on Breast Cancer and how FNA features can help classify the cancer as malignant or benign. Teamwork played a major role in the successful completion of our project.

Smruthi and Snigdha performed the EDA and visualisation in the first phase of the project, providing us with insights to the data and visualizations to have a better understanding of the data. Additionally worked on SVC model for prediction.

Vridhi worked on literature review and provided us with the background information on various works done on breast cancer predictions. Also helped in calculating the performance metrics for the different models used.

Yashi helped out with the literature review to provide background information for the already existing work for the problem and contributed towards working on providing different combinations of preprocessing techniques and by performing grid search to choose the best parameters for Logistic Regression, AdaBoost, Random Forest, KNN.

ACKNOWLEDGMENT

We wish to acknowledge Dr. Gowri Srinivasa, our Data Analytics Professor, for giving us the opportunity to work on this project which helped us enhance our knowledge. We would also like to thank the Teaching Assistants, Ms. Adithi Satish, Mr. Akhil Eppa, Mr. Lavitra Kshitij and Mr. Tejas Srinivasan who supported us throughout the course of this project. We would like to express our gratitude towards the Data Analytics faculty for providing us this platform to enrich our proficiency in this domain.

REFERENCES

[1] WHO Fact Sheet on Breast Cancer. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.
[2] 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).
[3] Abdulrahman Alalshekmubarak and Leslie S Smith. 2013. A novel approach combining recurrent neural network and support vector machines for time series classification. In Innovations in Information Technology (IIT), 2013 9th International Conference on. IEEE, 42–47
[4] 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).
[5] 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
[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).