IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

# Explainable ML and DL Models on Breast Cancer Data

**[1]Amit Kumar Dubey, [2]Abhilash Banerjee, [3]Abhishek Chakraborty, [4]Anitabha Das, [5]Arya Raj, [6]Apurba Paul**

[12345]UG Student, [6]Assistant Professor

[123456]Department of Computer Science of Engineering,

[123456]JIS College of Engineering, Kalyani, West Bengal, India.

***Abstract :*** In recent years, breast cancer has grown in importance. Women seem to be developing breast cancer at a considerably higher rate. If the illness is not detected at all, it has already become fatal, and in many cases, limb amputation is the only method to stop it if it is discovered too late. Therefore, a reliable indicator of this problem can aid in accurate diagnosis. Malignant breast cancer is a potentially life-threatening cancer that develops in the cells of the breast. It is most commonly found in women, but men can get it as well. Symptoms of malignant breast cancer can include a lump or thickening in the breast or armpit, changes in the size or shape of the breast, dimpling of the skin, redness or scaliness of the nipple, or discharge from the nipple. Benign breast cancer is a type of noncancerous breast growth that does not spread to other parts of the body. It is generally not life threatening and is usually treatable with surgery or radiation. Symptoms of benign breast cancer can include a lump or thickening in the breast, changes in the size or shape of the breast, dimpling of the skin, redness or scaliness of the nipple, or discharge from the nipple. This paper's major objective is to apply various machine learning classification algorithms to accurately forecast the target class and enhance it by evaluating the usefulness of specific aspects of the original Wisconsin Breast Cancer dataset (WDBC) for breast cancer diagnosis prediction. The highest performing algorithm was identified after classifiers were run on the dataset, and then useful dataset features were examined to boost performance even further. We employed the Naive Bayes, Support Vector Machine (SVM), Logistic Regression, Random Forest, CNN, RNN, FFNN algorithms in this paper. Performance metrics were employed to compare the results in this case, including accuracy, precision, recall, F-measure and ROC, AUC Curve. Among the algorithms utilised, convolutional neural network produced the best results based on the values of the performance indicators. On the same dataset, we also sought to optimise our suggested model and compare it to other cutting-edge methods given by other researchers.

***IndexTerms*** **- Breast Cancer, Dataset, CNN, RNN, Gaussian Naïve Bayes, Random Forest, SVM, Logistic Regression, Decision Tree, LIME, SHAP.**

## I. INTRODUCTION

Statistics published by the International Agency for Research on Cancer (IARC) in December 2020 show that breast cancer has now surpassed lung cancer as the most frequently diagnosed cancer in women globally. From a projected 10 million cases in 2000 to 19.3 million cases in 2020, the total number of cancer diagnoses has then doubled over the last two decades [1]. In the modern world, one in five people will have cancer at some point in their lives. According to projections, the number of persons receiving cancer diagnoses will rise even more in the upcoming years and will be approximately 50% higher in 2040.Cancer is the cause of more than one in six fatalities.

The most prevalent disease in women is breast cancer, according to the Centres for Disease Control and Prevention (CDC)Trusted Source. The large variations in breast cancer survival rates are caused by several variables. Both the type of cancer that women have and the stage of the disease when they obtain a diagnosis are the most crucial variables. Breast cells can evolve into cancer, which is called breast cancer. Usually, breast cancer develops in the ducts or lobules of the breast. Additionally, cancer can develop in your breast's fat tissue or fibrous connective tissue. In addition to often invading healthy breast tissue, unchecked cancer cells can also go to the lymph nodes beneath the arms. According to medical professionals, breast cancer was caused by breast cells that grew abnormally and then spread to the lymph nodes or other regions of the body. To prevent the effects of the following phase, it is vital to identify and stop the proliferation of these undesirable cells as soon as feasible. The first thing a doctor does after diagnosing a tumour is to determine if it is benign or malignant. because the two tumours have distinct treatment and preventative approaches. While malignant cells can travel to other areas of the body, benign cells are not carcinogenic and cannot do so. The issue with this condition is that there isn't a reliable diagnostic tool available to identify cancer in its earliest stages, allowing the patient to begin treatment as soon as possible and attempt to stop the spread of malignant cells or tumours.

Any sickness that is detected early enough to be treated with some degree of human effort. Most people miss their ailment before it progresses to chronicity. The death rate rises as a result all across the world. When it is discovered at an early stage, before it has spread to every region of the body, breast cancer is one of the diseases that may be cured.

The absence of prognostic models makes it challenging for doctors to develop a therapeutic strategy that could increase patient survival time. Therefore, it takes time to design the method that produces the least number of mistakes in order to enhance accuracy. There was a need for a computerised diagnostic system that employed Machine Learning technique since the existing methods to diagnose breast cancer, such as mammography, ultrasound, and biopsy, were time-consuming. This approach uses algorithms to more correctly and quickly identify cells while also assisting in the categorization of tumours.

Breast cancer, the most frequent malignancy is the second most common cause of death in women. It is a heterogeneous disease on the molecular level. Over the past 10–15 years, treatment concepts have evolved to treat this. It is very important to detect the disease and classify Malignant and Benign patients. One of the causes of increase in the number of breast cancer is change in food and lifestyle. It happens due to abnormal growth of fatty and fibrous tissues. The cells form tumours that can be seen in x-rays. The cause of breast cancer includes changes and mutations of DNA. Different stages are determined by the spreading of the cancer cells in tissues. These stages define how far a patient's cancer has proliferated. The spreading occurs when the cancer cells get carried to other parts through blood. There are different types of breast cancer. The common ones are ductal carcinoma in situ (DCIS) and invasive carcinoma. The side effects of this disease are – fatigue, headaches, pain and numbness, bone loss and osteoporosis. To prevent the spreading of the cancerous cell a patient has to undergo various treatments such as breast cancer surgery, chemotherapy, radioactive and endocrine. A breast cancer patient needs treatment on time so it is important to have the most accurate diagnosis result. A number of methods have been presented to get the accurate diagnosis. Machine learning can be applied for prediction because of the distinct report attribute in datasets. The technologies used till date cannot give the prediction full automatically. Hence, we have proposed a fully automatic method to classify and predict breast cancer based on the provided dataset using deep learning techniques. This technique is best recognised to predict and classify breast cancer from image dataset. The goal of the research is to identify and classify malignant and benign patients with the intention on how to parameterize the classification techniques to get high accuracy. The paper presents a comparison between the performance of eight algorithms among which convolutional neural network and Gaussain-Naive Bayes are the most influential data mining algorithms. Through the research we are looking after many datasets and check how farther machine learning algorithms can be used to characterise breast cancer. The aim is to reduce the error rate with more accuracy. A 10-fold cross-validation test is used in jupyter to evaluate the data and analyse data in terms of effectiveness and efficiency.

## II. METHODOLOGY

### 2.1 Objectives

The Model Learning aims to observe trends that may aid us in predicting Malignant or benign cancer. To achieve this use machine learning classification methods to fit a function that can predict the class of new input or test data.

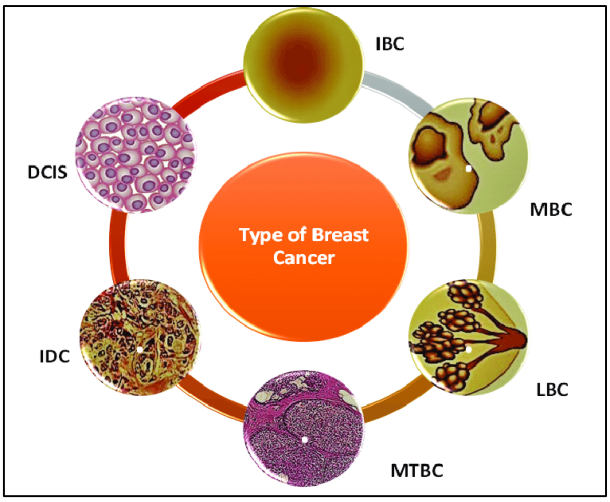


**Fig. 1** Type of Breast Cancer

### 2.2 Procedure

#### 2.2.1 Collecting data

- Machines initially learn from the data that you give them. It is of the utmost importance to collect reliable data so that your machine learning model can find the correct patterns. The quality of the data that you feed to the machine will determine how accurate your model is. If you have incorrect or outdated data, you will have wrong outcomes or predictions which are not relevant.
- The dataset includes ten features from each one of the cells in the sample. The Used Dataset includes 699 rows and 11 columns in which the last column represents classes of cancer cells and the rest ten columns are features of cancer cells.
  **Ten real-valued features are computed for each cell nucleus:**
  1. Sample code number.
  2. Clump Thickness.
  3. Uniformity of Cell Size.
  4. Uniformity of Cell Shape.
  5. Marginal Adhesion.
  6. Epithelial Cell Size.
  7. Bare Nuclei.
  8. Bland Chromatin.
  9. Normal Nucleoli.
  10. Mitoses.

**2.2.2 Preparing the data**

- Cleaning the data to remove unwanted data, missing values, rows, and columns, duplicate values, data type conversion, etc. You might even have to restructure the dataset and change the rows and columns or index of rows and columns.
- Splitting the cleaned data into two sets - a training set and a testing set. The training set is the set your model learns from. A testing set is used to check the accuracy of your model after training.
- Use the SciKit-Learn library to Split the dataset into training and test data using the train_test_split method. Divide dataset into train-dataset and test-dataset in ratio 80% by 20%. Divide train and test dataset into Input as X_train , X_test and Output as Y_train, Y_test. Now the input train-dataset contains 559 rows and 10 columns and the input test-dataset contains 132 rows and 10 columns. Whereas the output train-dataset contains 559 rows and 1 columns and output test dataset contain 132 rows and 1 column.
- Feature Scaling

Use StanadardScalar method from the SciKit-Learn library for feature scaling if the dataset contains features highly varying in range. Transform the dataset by pre-processing and Standard scaling. Since Dataset contains limited features therefore not use feature selection method here. Check the variation of output train dataset using Distplot.

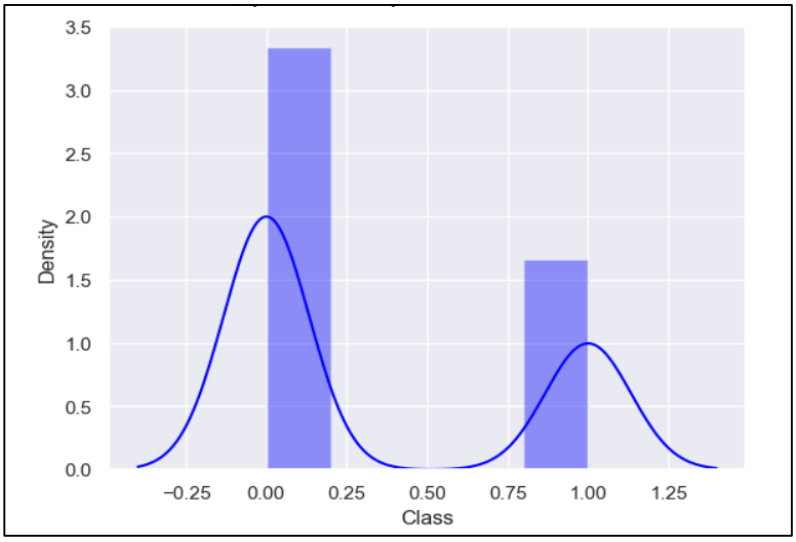


**Fig.2** Distplot

**2.2.3 Choosing a model**

A machine learning model determines the output you get after running a machine learning algorithm on the collected data. It is important to choose a model which is relevant to the task at hand.

**2.2.3.1 Logistic Regression**

Logistic regression is a classification algorithm. It is used to predict a binary outcome based on a set of independent variables. Use this algorithm by importing it from the Sklearn library of python. Then prepare the model by fitting into X_train and Y_train and predicting for X_test.

**2.2.3.2 Decision Tree**

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. This algorithm can be used for regression and classification problems — yet, is mostly used for classification problems. A decision tree follows a set of if-else conditions to visualise the data and classify it according to the conditions. Use this algorithm by importing it from the Sklearn library of python. Then prepare the model by fitting into X_train and Y_train and predicting for X_test.

**2.2.3.3 Random Forest**

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. Use this algorithm by importing it from the Sklearn library of python. Then prepare the model by fitting into X_train and Y_train and predicting for X_test.

**2.2.3.4 Gaussian Naive Bayes**

Naive Bayes is a generative model. (Gaussian) Naive Bayes assumes that each class follows a Gaussian distribution. The difference between QDA and (Gaussian) Naive Bayes is that Naive Bayes assumes independence of the features, which means the covariance matrices are diagonal matrices. Use this algorithm by importing it from the Sklearn library of python. Then prepare the model by fitting into X_train and Y_train and predicting for X_test.

**2.2.3.5 SVM**

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. Use this algorithm by importing it from the Sklearn library of python. Then prepare the model by fitting into X_train and Y_train and predicting for X_test.

#### 2.2.3.6 CNN

CNNs, or convolutional neural networks, are a sort of deep learning algorithm frequently used for computer vision applications including object identification and picture categorization. Due to its capacity to automatically learn hierarchical representations, CNNs have demonstrated to be quite efficient in analysing visual data. Convolutional, Activation, Pooling, Fully Connected, and Output layers make up the CNN algorithm. Fit the model into X_train and Y_train, and then make predictions for X_test.

#### 2.2.3.7 FFNN

FFNN or Feed Forward Neural Network is a type of neural network which is used to a sort of artificial neural network in which data only travels in one way, from the input layer to the output layer, without looping around or repeating itself. It is made up of neurons and functions like as activation, forward propagation, learning, loss, and inference.

#### 2.2.3.8 RNN

RNN or Recurrent Neural Network is a kind of artificial neural network made to process time series, speech, text, and video, among other sequential data types. RNNs include loops in their internal structure that enable them to process input sequences in a recursive fashion, in contrast to typical neural networks, which only process input data in a single direction. This implies that RNNs can draw on their recollection of previous inputs to guide how they process the inputs they receive today.

#### 2.2.3.9 LSTM

The recurrent neural network (RNN) variant known as LSTM, or "Long Short-Term Memory," was created to address the issue of disappearing gradients in conventional RNNs.

Compared to conventional RNNs, LSTMs have a more intricate internal structure made up of numerous gates that control the information flow across the network. An input gate, an output gate, and a forget gate are some of these gates. The output gate regulates how much of the current state is used as output, the forget gate regulates how much of the prior state is forgotten, and the input gate regulates how much new input is permitted into the network.

### 2.2.4 Training a model

Training is the most important step in machine learning. In training, you pass the prepared data to your machine learning model to find patterns and make predictions. Over time, with training, the model gets better at predicting.

### 2.2.5 Evaluating a model

Confusion matrix- A confusion matrix is a table that is used to define the performance of a classification algorithm. A confusion matrix visualizes and summarizes the performance of a classification algorithm.

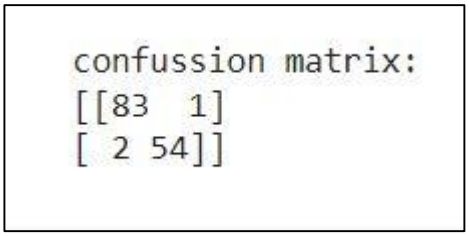


**Fig.3** confusion matrix

After training your model, you have to check to see how it's performing. This is done by testing the performance of the model on previously unseen data. The unseen data used is the testing set that you split our data into earlier. It shows how our model will perform and its speed.

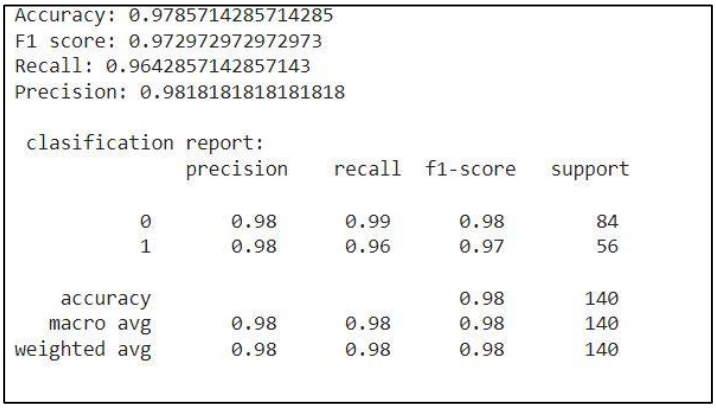


**Fig.4** accuracy, f1 score, recall and precision

### 2.2.6 Improving the model

One of the biggest challenges in all of these ML and DL projects in different industries is model improvement because the number of AI use cases has been increasing exponentially with the rapid development of new algorithms, cheaper compute, and greater availability of data.

**2.2.6.1 Cross validation**

k-fold cross validation

```
[0]logistic regression accuracy: 0.9714285714285714
[1]Decision tree accuracy: 0.9285714285714286
[2]Random forest accuracy: 0.9785714285714285
[3]Gaussian Naive Bayes accuracy: 0.9571428571428572
[4]SVM accuracy: 0.9714285714285714
```

**fig.5** without k fold

```
[0]logistic regression accuracy: 0.9606818181818182
[1]Decision tree accuracy: 0.9409740259740259
[2]Random forest accuracy: 0.9606493506493505
[3]Gaussian Naive Bayes accuracy: 0.9624675324675325
[4]SVM accuracy: 0.9606818181818182
```

**Fig.6** with k-fold

**2.2.6.2 Hyperparameter tunning**

The process of choosing the best settings for a machine learning algorithm's or model's hyperparameters is known as hyperparameter tuning. Hyperparameters are the options or configurations that are chosen by the user before to the learning process and are not learnt from the data. They manage the model's performance and behaviour while it is being trained.

The common techniques used in hyperparameter tunning is Grid Search

**Grid search**

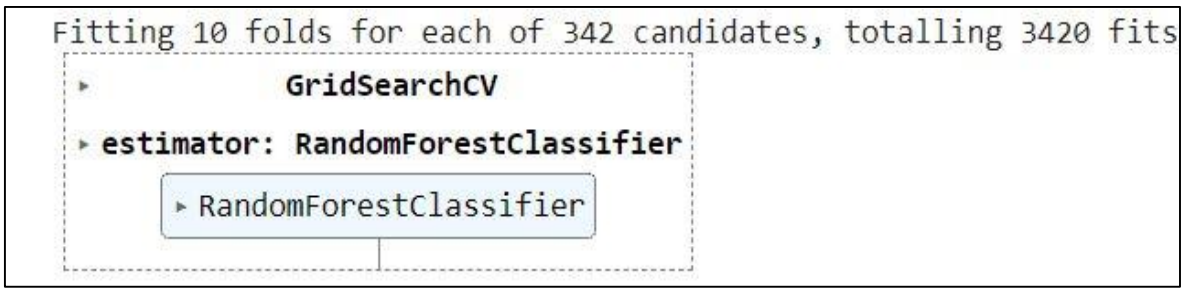


**Fig 7** grid search

**Bagging & Boosting**

**Table 1** bagging and boosting

| | |
|---|---|
| Decision Tree: | 0.964221824686941 |
| Random Forest | 0.9731663685152058, |
| Gaussian Naïve Bayes | 0.962432915921288, |
| Support Vector Classifier | 0.964221824686941 |
| Logistic Regression | 0.964221824686941 |

## III. Final Accuracy of Models

**Table 2** accuracy of models

| Algorithms | Accuracy |
|---|---|
| Logistic regression | 96% |
| Decision Tree | 94.1% |
| Random Forest | 96% |
| SVM | 96.068% |
| Gaussain Naive Bayes | 96.2% |
| CNN | 99.8% |
| FFNN | 67.6% |
| RNN(LSTM) | 87.3% |

### 3.1 Python Library used during Training of Model

1. Numpy
2. Pandas
3. Matplotlib
4. Seaborn
5. Sklearn
6. Tensorflow
7. Keras
8. Explainable AI

### 3.2 Visualizing the models using Explainable AI(XAI)

Explainable AI refers to methods and techniques in the application of artificial intelligence technology (AI) such that the results of the solution can be understood by human experts. It contrasts with the concept of the “black box” in machine learning where even their designers cannot explain why the AI arrived at a specific decision. XAI is an implementation of the social right to explanation

### 3.3 Using LIME

Local Interpretable Model Agnostic Explanation is referred to as LIME. LIME's main advantages are how easy it is to use and understand. LIME is broad, yet its core idea is relatively simple and basic. Let's first examine what the name actually means: LIME has a characteristic known as model agnosticism that enables it to justify any particular supervised learning model by treating it as a standalone "black box." Therefore, nearly every model now in use may be supported by LIME. Local explanations are reasons offered by LIME that are precise in the vicinity of the observation or sample being explained.

Table 3. LIME explanations in dataset

| Lime using ML Model | Decision Tree | Random Forest | Logistic Regression | Gaussian NB | SVC |
|---|---|---|---|---|---|
| **Prediction Probabilities for Malignant** | 1.00 | 0.98 | 0.98 | 1.00 | 0.97 |
| **Bare Nuclei** | 0.41 | 0.32 | 0.32 | 0.28 | 0.33 |
| **Marginal Adhesion** | 0.13 | 0.26 | 0.27 | 0.24 | 0.26 |
| **Uniformity of Cell Shape** | 0.09 | 0.21 | 0.20 | 0.23 | 0.14 |
| **Mitoses** | -0.06 | -0.22 | -0.22 | -0.19 | -0.31 |
| **Normal Nucleoli** | -0.05 | 0.00 | 0.02 | -0.05 | 0.02 |
| **Uniformity of cell Size** | -0.04 | -0.00 | 0.01 | -0.04 | -0.01 |
| **Single Epithelial cell Size** | 0.03 | 0.01 | -0.01 | -0.03 | -0.00 |
| **Clump Thickness** | 0.03 | 0.04 | 0.03 | -0.00 | 0.04 |
| **Bland Chromatin** | 0.00 | 0.01 | 0.02 | -0.02 | 0.04 |

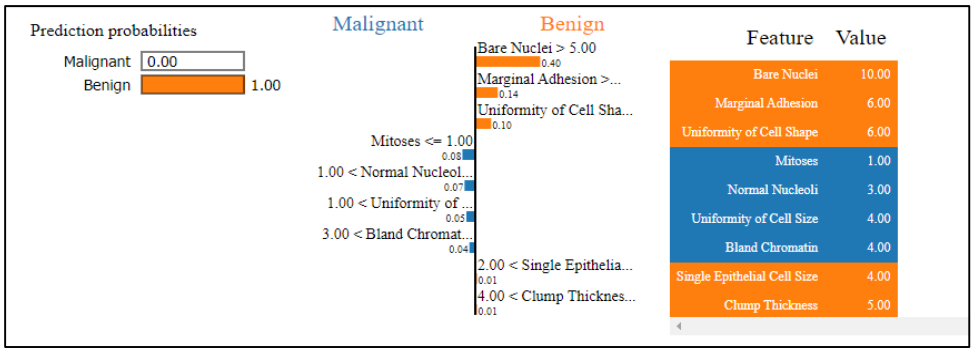


**Fig. 8** Explaining Decision Tree Using LIM

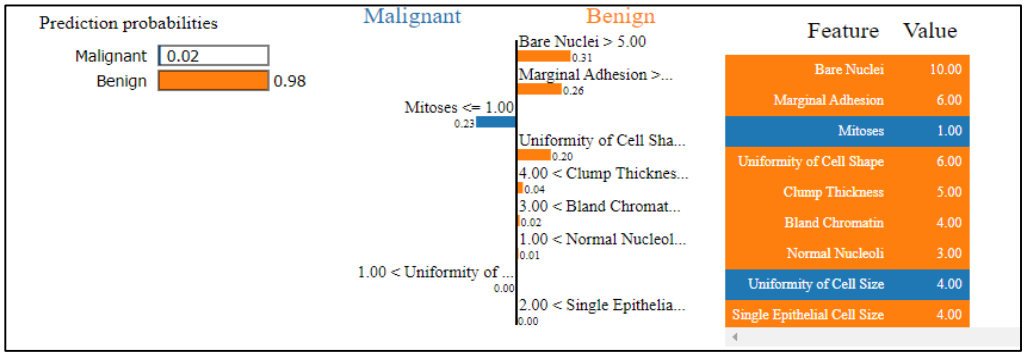


**Fig. 9** Random Forest Tree Using LIME

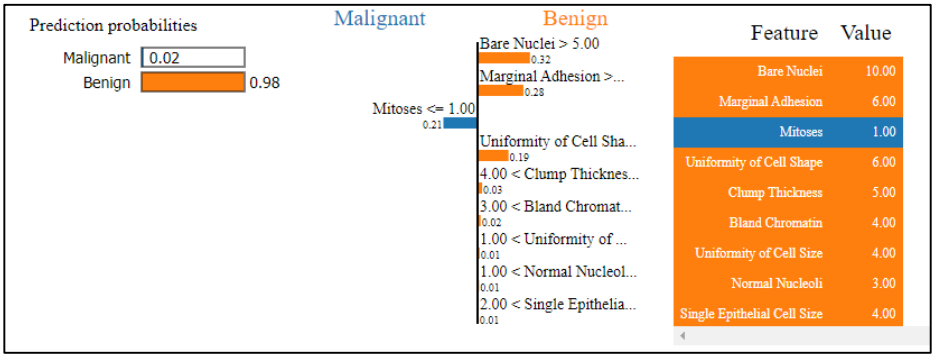


**Fig. 10** Logistic Regression Tree Using LIME

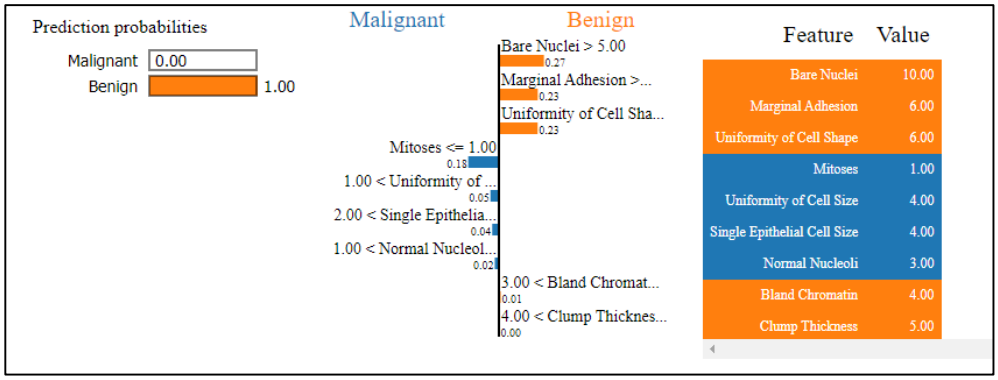


**Fig. 11** Gaussian-Naive Bayes Tree Using LIME

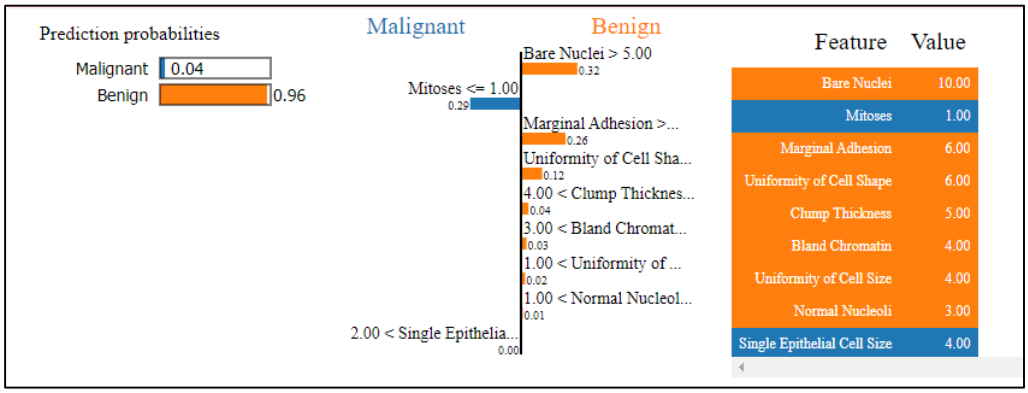


**Fig. 12** Support Vector Machine Using LIME

### 3.4 Using SHAP

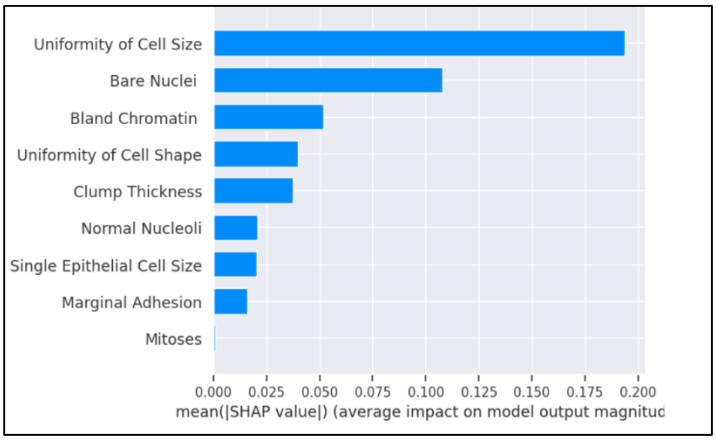


**Fig. 13** summary plot using SHAP

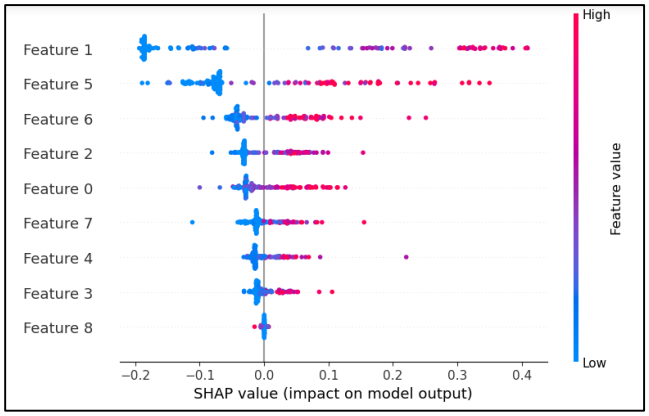


**Fig. 14** beeswarm plot using SHAP

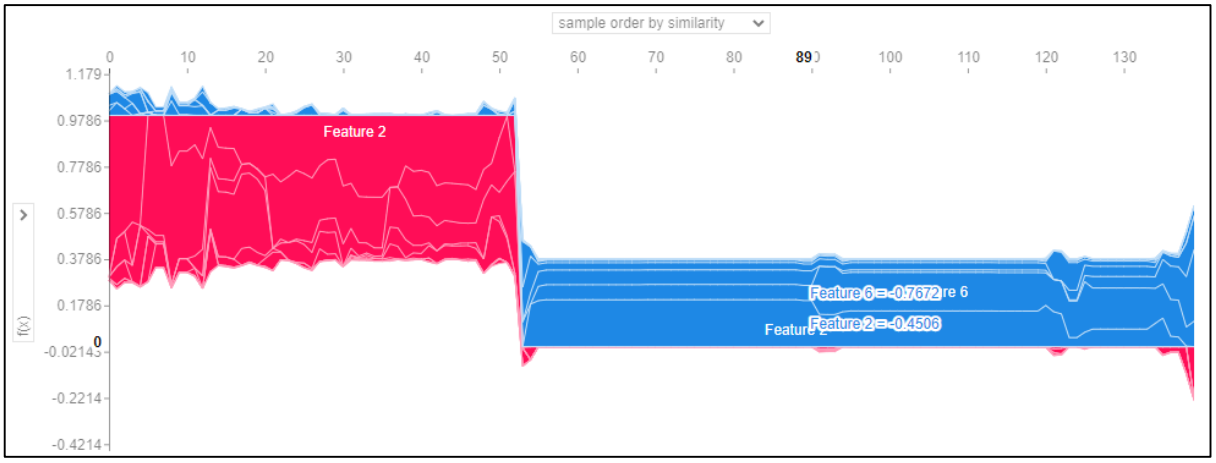


**Fig. 15** force plot using SHAP for Decision Tree

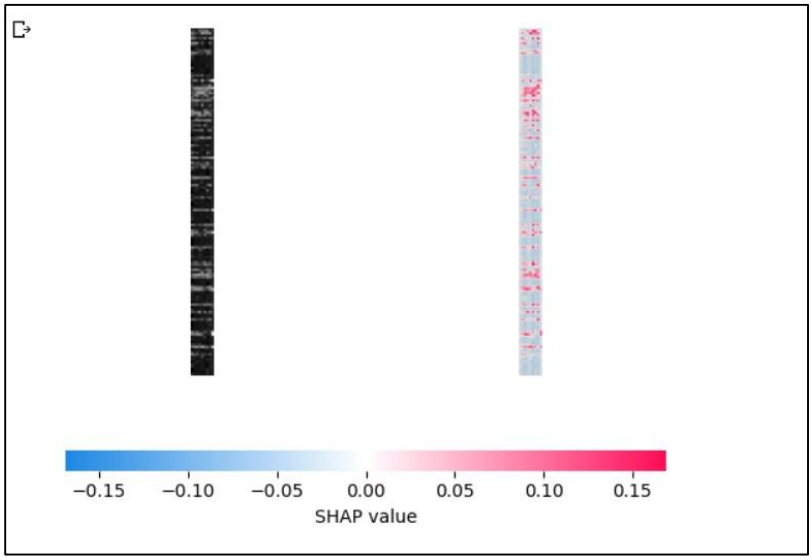


**Fig. 16** Image plot of CNN using shap

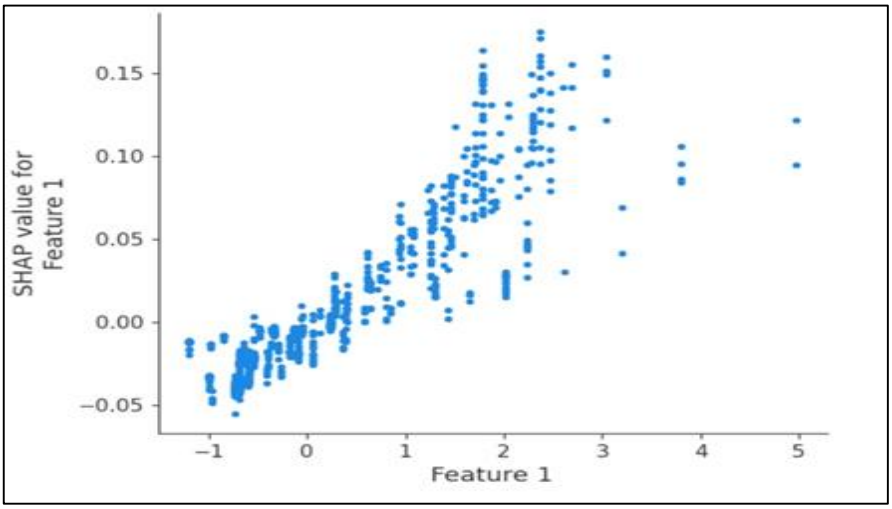


**Fig. 17** Dependence plot of CNN using shap

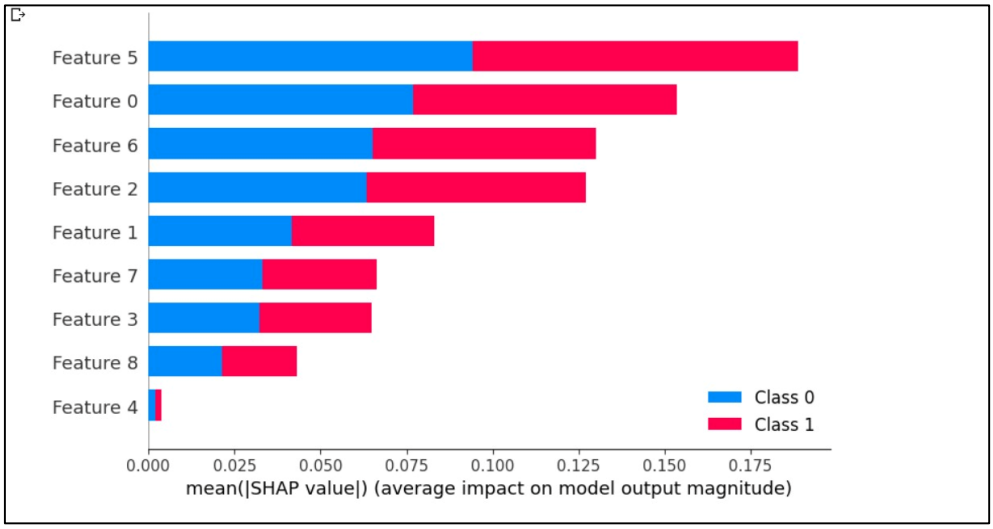


**Fig. 18** Summary plot of FFNN using shap

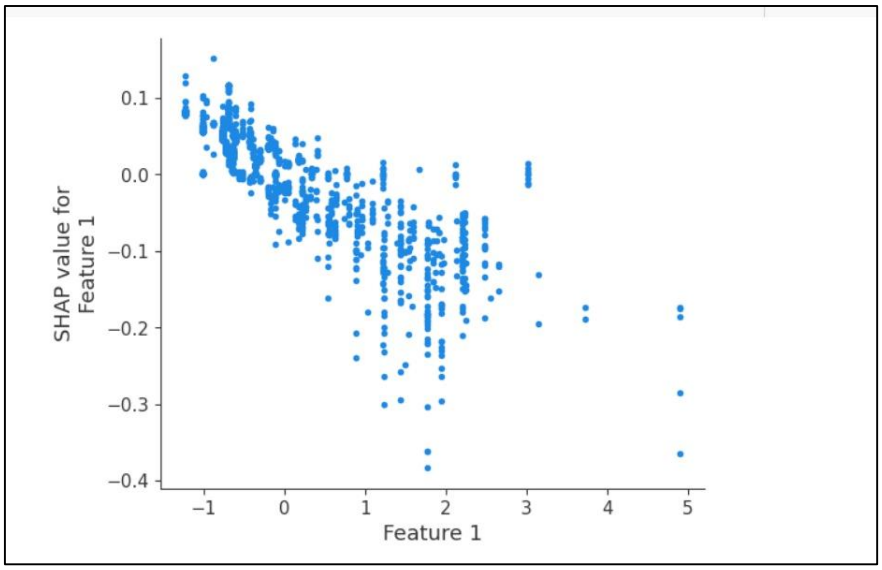


**Fig. 19** Dependence plot of FFNN using shap

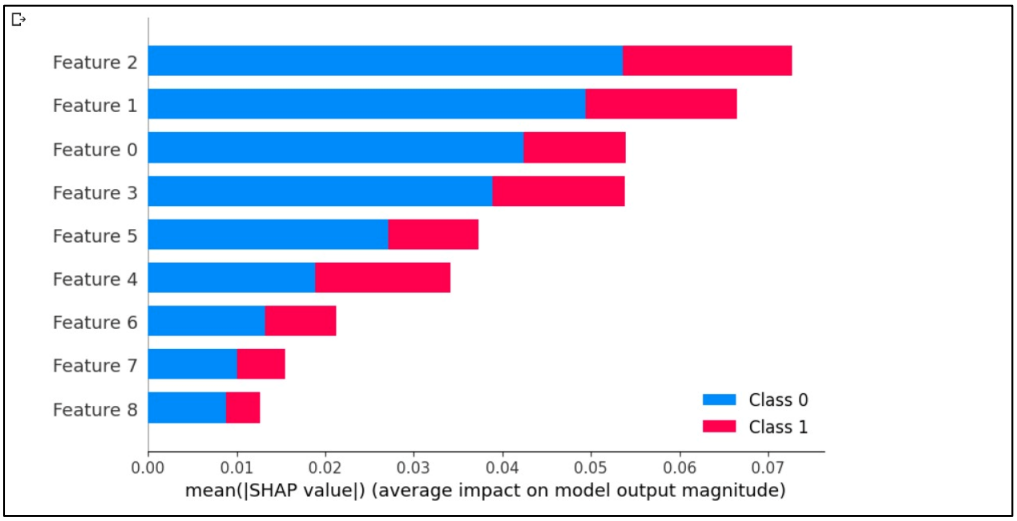


**Fig. 20** Summary plot of RNN using shap

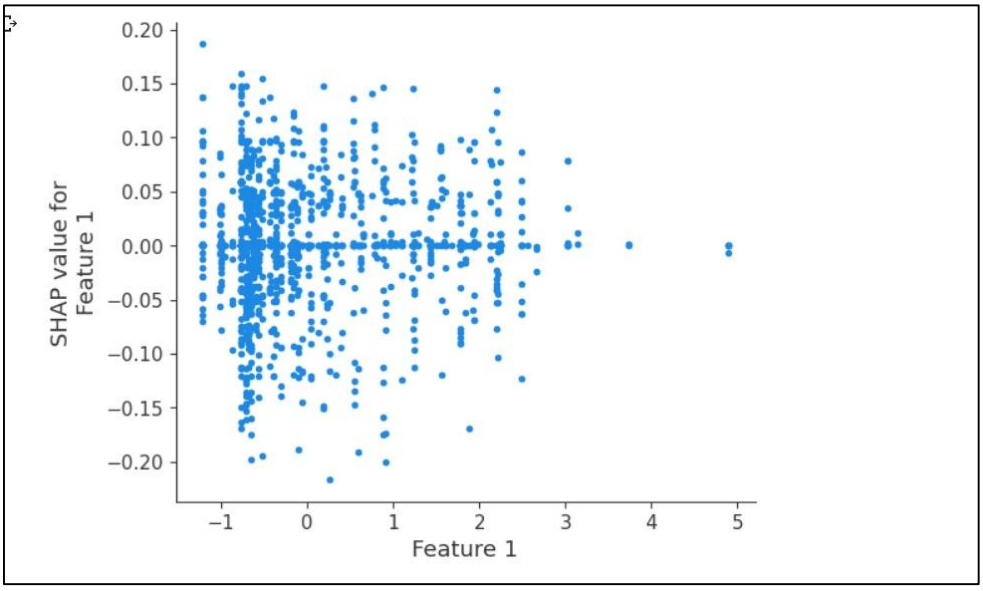


**Fig. 21** Dependence plot of RNN using shap

## IV. RESULT ANALYSIS

We have tested different types of algorithms for this dataset until now. Each algorithm is suitable for different conditions and different types of datasets. We found that after hyperparameter tuning these eight algorithms Random Forest Classifier, Decision Tree Classifier, Logistic Regression, Gaussian Naive Bayes and Support Vector Classifier, CNN, FFN and LSTM gives the accuracy on this dataset are 98.1%, 96%, 97.6%, 95.2%, 97.2%, 99.8%, 67.6%,87.3%. After analysing these algorithms, we came to know that CNN is the most suitable algorithm for this dataset of Breast Cancer, as the accuracy of this algorithm is the highest (99.8%) among other algorithms. So, we found out that CNN is the most suitable algorithm for this dataset. For the prediction of breast cancer through machine learning techniques, the major challenge is the availability of datasets. Each algorithm requires a huge amount of training data for its computational measurements to get the best output.

## V. CONCLUSION

In this paper we have reviewed different types of machine learning and deep learning algorithms for predicting breast cancer. And we have also explained the algorithms using the Explainable AI. In explainable AI we have used SHAP and LIME. Our goal is to find out the most suitable algorithm that can predict the occurrences of breast cancer more effectively. The review of this paper started from the types of breast cancer, symptoms and causes of breast cancer. After that different types of machine learning algorithms are used to know which technique will be best for this prediction. And finally, we got the result. Breast Cancer if found at early stage will save the lives of thousands of people. By using machine learning we can predict the type of the breast cancer benign or malignant as per the patent's biopsy report. These machine learning techniques can be used for medical research. It can reduce human error and manual mistakes.

## REFERENCES

**[1]** Amarta Kundu, Pabitra Kumar Bhunia, Poulami Mondal, Monalisa De, Sumanta Chatterjee, "MACHINE LEARNING-BASED HOME PRICE PREDICTION", IJRAR - International Journal of Research and Analytical Reviews (IJRAR), E-ISSN 2348-1269, P- ISSN 2349-5138, Volume.9, Issue 4, Page No pp.335-339, October 2022, Available at: http://www.ijrar.org/IJRAR22D1732.pdf

**[2]** A. Rovshenov and S. Peker, "Performance Comparison of Different Machine Learning Techniques for Early Prediction of Breast Cancer using Wisconsin Breast Cancer Dataset," 2022 3rd International Informatics and Software Engineering Conference (IISEC), Ankara, Turkey, 2022, pp. 1-6, doi: 10.1109/IISEC56263.2022.9998248.

**[3]** X. Zhang and Y. Sun, "Breast cancer risk prediction model based on C5.0 algorithm for postmenopausal women," 2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), Jinan, China, 2018, pp. 321-325, doi: 10.1109/SPAC46244.2018.8965528.

**[4]** G. Shanmugasundaram, S. Balaji, R. Saravanan, V. Malarselvam and S. Yazhini, "SYSTEMATIC ANALYSIS ON BREAST CANCER PREDICTION," 2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA), Pondicherry, India, 2018, pp. 1-5, doi: 10.1109/ICSCAN.2018.8541239.

**[5]** J. Aditya, "Optimized Ensemble Prediction Model for Breast Cancer," 2021 International Conference on Intelligent Technology, System and Service for Internet of Everything (ITSS-IoE), Sana'a, Yemen, 2021, pp. 1-4, doi: 10.1109/ITSS-IoE53029.2021.9615269.

**[6]** S. Nathiya and J. Sumitha, "A Comparative Study on Breast Cancer Prediction using Optimized Algorithms," 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2021, pp. 1401-1405, doi: 10.1109/ICOSEC51865.2021.9591787.

**[7]** X. Feng et al., "Accurate Prediction of Neoadjuvant Chemotherapy Pathological Complete Remission (pCR) for the Four Sub-Types of Breast Cancer," in IEEE Access, vol. 7, pp. 134697-134706, 2019, doi: 10.1109/ACCESS.2019.2941543.

**[8]** D. Yifan, L. Jialin and F. Boxi, "Forecast Model of Breast Cancer Diagnosis Based on RF-AdaBoost," 2021 International Conference on Communications, Information System and Computer Engineering (CISCE), Beijing, China, 2021, pp. 716-719, doi: 10.1109/CISCE52179.2021.9445847.