

BREAST CANCER CLASSIFICATION

by

P. Sai Gangadhar Kumar

1. ABSTRACT

Around the world, a wide range of applications use machine learning and deep learning. This is also true in the healthcare sector. There are two goals for this essay. We use a variety of deep learning models to classify images of breast cancer histology as the initial goal. This study identified the most precise algorithms for categorizing breast cancer histopathology image databases. The different accuracy scores for the deep learning models on the same database demonstrated the importance of other aspects including Data Preparation, Data description, data visualization, Data Analysis. The models like ResNet, GoogleNet, VGGNet, AlexNET, and EfficientNET had been identified that each of them provided different results for the Image database. we analyse both existing and proposed classifiers. We have shown the comparison and evaluation of the accuracy of these five implementations. Results indicated that VGGNET has shown the highest accuracy of 83%. The second goal is to classify breast cancer using various machine learning techniques. To do this, we used a tabular dataset which was developed with 31 features. Nowadays, machine learning (ML) techniques are widely used to the problem of breast cancer classification. High classification accuracy and efficient diagnostic abilities are provided. In this project, we analyse classifiers such as decision trees, logistic regression, XGBoost, RandomForest, K-Nearest Neighbour (KNN), and Back propagation and Multi-layer perceptron. We then propose an ensemble classifier that performs hybrid classification by using both strong and weak classifiers because it can have a large number of training and validation samples. So, in order to improve classification and accuracy, we analyze both existing and proposed classifiers. We have shown the comparison and evaluation of the accuracy of these five implementations. Results indicated that both decision trees and random forests had accuracy levels of over 90%. As medical facts are not ready in a greatly greater size we use facts Data Visualization to make use of the ready facts more with small amounts of support. True medical facts are frequently incomplete, inconsistent, lacking, absent from particular behaviors or tendencies, and are likely to contain numerous inaccuracies.

Keywords — Machine Learning, Deep Learning, Breast Cancer, Decision Tree, Logistic Regression, Random Forest, XGBoost, KNN, Artificial Neural Networks, Back Propagation, Multi-Layer Perceptron, VGGNET, AlexNET, RESNET, GoogleNET, EfficientNET.

ACKNOWLEDGEMENT

First and foremost, we would want to express our gratitude to Dr. Ravi Kant Kumar sir, assistant professor in the computer science and engineering department at SRM University AP. A special word of gratitude to Dr. Jathindra Kumar Dash, Head of Department, Computer Science and Engineering Department, SRM University AP, for this continued guidance and support for our project work.

S.NO	CONTENT	PAGES
1.	ABSTARCT	2
	ACKNOWLEDGEMENT	3
	TABLE OF CONTENTS	4
2.	INTRODUCTION	5
3.	LITERATURE REVIEW / RELATED WORK	5 to 11
4.	PROPOSED METHOD (4A. FLOWCHART)	11 to 14
5.	DATA 5A. Data Preparation 5B. Data Description 5C. Data Analysis / Data Visualization 5D. Visualization plots	15 to 22
6.	TESTING AND TRAINING OF MODELS	22 to 44
7.	RESULY ANALYSIS	44 to 45
8.	CONCLUSION	45
9.	FUTURE SCOPE	45 and 46
10.	REFERENCES	46 to 49

2. INTRODUCTION

One of the most common types of cancer among women worldwide is breast cancer. In order to improve patient outcomes and survival rates, early detection and precise classification of breast cancer are essential. Machine learning and deep learning algorithms have become effective tools for classifying breast cancer as a result of technological breakthroughs. Algorithms for machine learning and deep learning can analyse vast volumes of data and find patterns or links that might be difficult for a human to see. These algorithms can be trained on labelled data, which consists of information on breast cancer cases as input, such as patient records or medical imaging, coupled with matching class labels indicating whether the case is benign or malignant. Using machine learning and deep learning algorithms, the goal of breast cancer classification is to create models that can precisely predict the likelihood of a breast cancer case being malignant or benign. These algorithms may learn complicated patterns and generate predictions based on such patterns by using numerous features collected from medical images or patient data. Algorithms for machine learning and deep learning are advantageous for classifying breast cancer because they can handle vast volumes of data and automatically learn from it. As they are exposed to additional data over time, these algorithms can also change and become more effective. Because of this, they serve as useful instruments for helping medical practitioners identify patients more quickly and accurately. In this study, we investigate the classification of breast cancer using machine learning and deep learning techniques. We investigate the performance of various algorithms in accurately classifying benign and malignant breast cancer patients, including support vector machines, random forests, convolutional neural networks, and recurrent neural networks. The findings of this study may aid in the creation of more accurate and effective diagnostic tools for breast cancer, which would eventually enhance patient treatment and outcomes.

3. LITERATURE REVIEW / RELATED WORK

[Amrane.M, Oukid.S, Gagaoua I,Ensari T.\[1\]](#)Breast cancer is caused by a combination of variables, including family history, obesity, hormones, radiation therapy, and even reproductive factors. Every year, one million women are newly diagnosed with breast cancer; according to a

World Health Organisation report, half of them will die because doctors frequently find the cancer late. Breast cancer is created by a mistake or mutation in a single cell, which can be shut down by the system or causes uncontrolled cell division. Malignant tumours spread to neighbouring cells, which can cause metastasis or spread to other areas, however benign masses cannot spread to other tissues, therefore their expansion is limited to the benign mass. Because of the absence of symptoms at the start of the disease, detecting BC may be difficult. However, with certain clinical testing, an appropriate diagnosis should be able to discriminate between benign and malignant tumours. A good detection has a low rate of false positives (FP) and false negatives (FN)[1]. Machine learning is a set of tools used to create and evaluate algorithms for prediction, pattern recognition, and classification. ML is composed of four steps: Data collection, model selection, model training, and model testing. The link between BC and ML is not new; it has been used for decades to classify tumours and other malignancies, predict sequences of cancer-causing genes, and predict prognosis. The classification's goal is to place each observation in the appropriate category. We employed two machine learning classifiers in this study: the Nave Bayesian Classifier and the k-nearest neighbour. The goal is to establish whether a patient's tumour is benign or malignant. In this study, we tailor two machine learning techniques for breast cancer categorization. The goal of this paper is to build effective machine learning algorithms for cancer classification in a data set employing two classifiers. Each classifier's performance will be evaluated in terms of accuracy, training procedure, and testing method. The study conducted aimed to classify breast cancer using machine learning techniques. The authors used a dataset of 569 breast cancer cases and applied various machine learning algorithms, including decision tree, k-nearest neighbor, and support vector machine, to classify the cases as either malignant or benign. The results showed that the support vector machine algorithm achieved the highest accuracy of 97.54%, followed by the decision tree algorithm with an accuracy of 94.74% [1]. The authors concluded that machine learning techniques can be effective in classifying breast cancer and can potentially assist in the diagnosis and treatment of the disease. However, the study was limited by the small sample size and the use of only one dataset. Further research is needed to validate the findings and to explore the potential of machine learning in breast cancer classification.

[2] To distinguish between benign and malignant conditions, the SVM classifier uses features taken from segmented pictures. Circularity range ratio, regularity variance, and perimeter are

among the features that were extracted. An anisotropic filter was used to eliminate noise from the 323 ultrasound breast pictures produced by the CAD, of which 216 included benign lesions and 107 contained malignant lesions. The CAD uses an automatic and semiautomatic process. When the semi automatic approach was used, the proposed method produced accuracy, sensitivity, and specificity results of 95.98%, 97.20%, and 95.37%, respectively. Likewise, when using fully automatic segmentation, the method has an accuracy and specificity of 95.67% and 95.83%, respectively, and a sensitivity of 95.33%.

Burak Akbugday.[3]

According to US cancer is the second leading cause of death in United States

Breast Cancer is 10th type of cancer that leads to death and it's more commonly seen in women. Over the last two decades the use of machine learning increased to several fields including medicine. The main advantage of Machine Learning is it can be defined as computational strategy which can be used to determine optimal solutions to a given problem. Now it's possible to analyze medical data, which is so difficult to analyze manually with the help of machine learning algorithms. The data set we are using for this research is Breast Cancer Wisconsin as this data set is public it is widely used for many machine learning studies. Many studies are published using this dataset. This data set contains 699 instances,10 attributes and a class attribute [3].

The authors of this research have shown Support Vector Machine-Radial basis Function (SVM-RBF) kernel is the most accurate classifier with 96.84% accuracy another researcher have made a similar comparison and found that with 96.99% accuracy, SVM classifier was the best. Breast cancer Wisconsin dataset has been used to investigate the effectiveness of Naive Bayes (NB), K-Nearest neighbors and SVM learning algorithms.

Burak Akbugday, Epimack Michael, He Ma, Hong Li, and Shou liang Qi. [4]

To distinguish between benign and malignant conditions, the SVM classifier uses features taken from segmented pictures. Circularity range ratio, regularity variance, and perimeter are among the features that were extracted. An anisotropic filter was used to eliminate noise from the 323 ultrasound breast pictures produced by the CAD, of which 216 included benign lesions and 107 contained malignant lesions. The CAD uses an automatic and semiautomatic process. When the semi-automatic approach was used, the proposed method produced accuracy, sensitivity, and

specificity results of 95.98%, 97.20%, and 95.37%, respectively. Likewise, when using fully automatic segmentation, the method has an accuracy and specificity of 95.67% and 95.83%, respectively, and a sensitivity of 95.33%.

Based on the morphological characteristics of breast ultrasound, Zeebaree et al. [4] created a CAD (computer aided design) that uses ML and region-growing segmentation. To extract features from the ROI, the technique uses a hybrid model. Instead of just one feature, the features include 7 moments, and HOG (Histogram of Oriented Gradients). We used 250 ultrasound scans, of which 100 showed benign lesions and 150 showed cancerous lesions. The ANN, which is used to classify ultrasound images, has an accuracy rate of 93.1% for cancerous lesions and 90.4% for benign lesions.

A technique for identifying and categorizing breast cancer using B-mode and elastography pictures was put forth by Adel et al. [4]. The suggested method uses 82 ultrasound images in total, of which 26 were benign lesions and 56 were malignant lesions. Geometrical and textural features became the basis for the extracted features from ROI (Region of Interest). The mean, standard deviation, area, perimeter, width-to-height ratio, contrast-to-noise ratio, and signal-to-noise ratio are among the 33 features retrieved from B-mode and elastography images. The SVM was used, and it had a 94.12% accuracy rate.

An XGBoost classifier for breast cancer classification was proposed by Chang et al. [4]. From the Chung Shan Medical University Hospital, Jen-Ai Hospital, and Far Eastern Memorial Hospital, a total of 2964 breast cancer samples have been collected. The results showed that the single XGBoost approach had a testing accuracy of 94.00%, which is very high.

Chugh, G., Kumar, S. and Singh, N [5]

The authors of the article "Survey on Machine Learning and Deep Learning Applications in Breast Cancer Diagnosis" by Chugh, Kumar, and Singh (2021) give a thorough overview of the application of machine learning (ML) and deep learning (DL) methods in the detection of breast cancer.

The authors begin by outlining the value of an accurate and prompt diagnosis of breast cancer and the drawbacks of conventional approaches. After that, they explore the several ML and DL techniques used in this field, such as support vector machines, random forests, artificial neural

networks, convolutional neural networks, and recurrent neural networks. Each algorithm's advantages and disadvantages are explained along with its performance metrics, learning curve, and interpretability. In ML and DL applications, data preprocessing and acquisition are essential steps [5]. The study examines publicly accessible databases like the Digital Database for Screening Mammography (DDSM) and the Digital Imaging and Communications in Medicine (DICOM) standard and emphasizes the importance of high-quality and diversified datasets. The authors also discuss difficulties with data preprocessing, including feature extraction, picture improvement.

Accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC) are among the evaluation metrics investigated to evaluate the performance of ML and DL models in the diagnosis of breast cancer. To ensure that models can be applied across different datasets, the authors emphasize the significance of model validation. Additionally, the difficulties associated with interpretability, legal compliance, and ethical considerations are addressed in relation to the prospective integration of ML and DL algorithms into clinical practice.

For a valuable overview of ML and DL applications in breast cancer diagnosis, see Chugh, Kumar, and Singh (2021). The page provides information on various algorithms, methods for gathering and processing data, metrics for measuring performance, and the potential application of these methods in clinical settings. The authors advance ML and DL-based breast cancer diagnosis by outlining the field's advantages, drawbacks, and potential prospects for the future.

Jannesari, M., Habib Zadeh, M., Aboulkheyr, H., Khosravi, P., Elemento, O., Totonchi, M. and Hajirasouliha, I [6]

Millions of women worldwide are affected by breast cancer, which is a serious global health issue. Effective treatment and better patient outcomes depend critically on the precise and early identification of breast cancer. Deep learning algorithms have recently demonstrated significant promise for automating the evaluation of histopathology images, assisting in the detection and categorization of breast cancer. The work "Breast Cancer Histopathological Image Classification: A Deep Learning Approach" by Jannesari et al., which was published in the 2018

IEEE International Conference on Bioinformatics and Biomedicine (BIBM), is summarized and critically analyzed in this review of related literature.

A deep learning-based method for categorizing breast cancer histopathology pictures is suggested by Jannesari et al. For the purpose of extracting pertinent features from the images and categorizing them into several histological classifications, the authors use a modified version of the convolutional neural network (CNN) architecture known as the VGGNet [6]. The study's dataset includes a sizable number of histopathological pictures that represent different types and subtypes of breast cancer. The effectiveness of the suggested strategy is confirmed by the performance review. The outcomes reveal improved classification accuracy when compared to baseline models, highlighting the promise of deep learning in the detection of breast cancer. However, additional testing on outside datasets and clinical trials would increase the suggested model's robustness and generalizability.

For the categorization of breast cancer histological pictures, Jannesari et al.'s work "Breast Cancer Histopathological Image Classification: A Deep Learning Approach" offers an enticing deep learning-based approach. The study adds to the body of work on the use of deep learning and artificial intelligence to the examination of medical images. The results show how deep learning models have the potential to increase the precision and effectiveness of breast cancer diagnosis. Future research should concentrate on validating external datasets and examining the therapeutic usefulness of the suggested approach.

Liew, X.Y., Hameed, N. and Clos, J., 2021 [7]

Breast cancer is a major public health issue that calls for precise and effective categorization techniques for early identification and treatment. In their study titled "An investigation of XGBoost-based algorithm for breast cancer classification," Liew et al. (2021) go into detail about the use of the XGBoost algorithm in this field. Their goal is to investigate how well XGBoost performs in accurately classifying breast cancer patients, maybe offering healthcare providers a useful tool.

The efficacy of the XGBoost algorithm in classifying breast cancer is thoroughly examined by Liew et al. in 2021. Their study makes use of a dataset that includes the clinical and histological characteristics of patients with breast cancer. The authors do a thorough study of XGBoost's

performance against that of other well-known classification algorithms, including Random Forest and Support Vector Machines, to establish its supremacy in terms of accuracy, precision, recall, and F1-score [7].

The findings published by Liew et al. (2021) show that XGBoost performs admirably in the categorization of breast cancer. The system surpasses conventional machine learning methods with a phenomenal accuracy rate of 95%. Additionally, XGBoost displays improved precision, recall, and F1-score, demonstrating its effectiveness in distinguishing between benign and malignant cases. This work highlights the potential of XGBoost to improve the precision and effectiveness of breast cancer classification systems.

The results of this study add to the corpus of information on machine learning algorithms used to categorize breast cancer. Making use of cutting-edge methods like XGBoost can increase the precision of diagnosis and treatment planning. The importance of XGBoost as a tool for healthcare practitioners is demonstrated by its capacity to handle high-dimensional and complex data related to breast cancer patients.

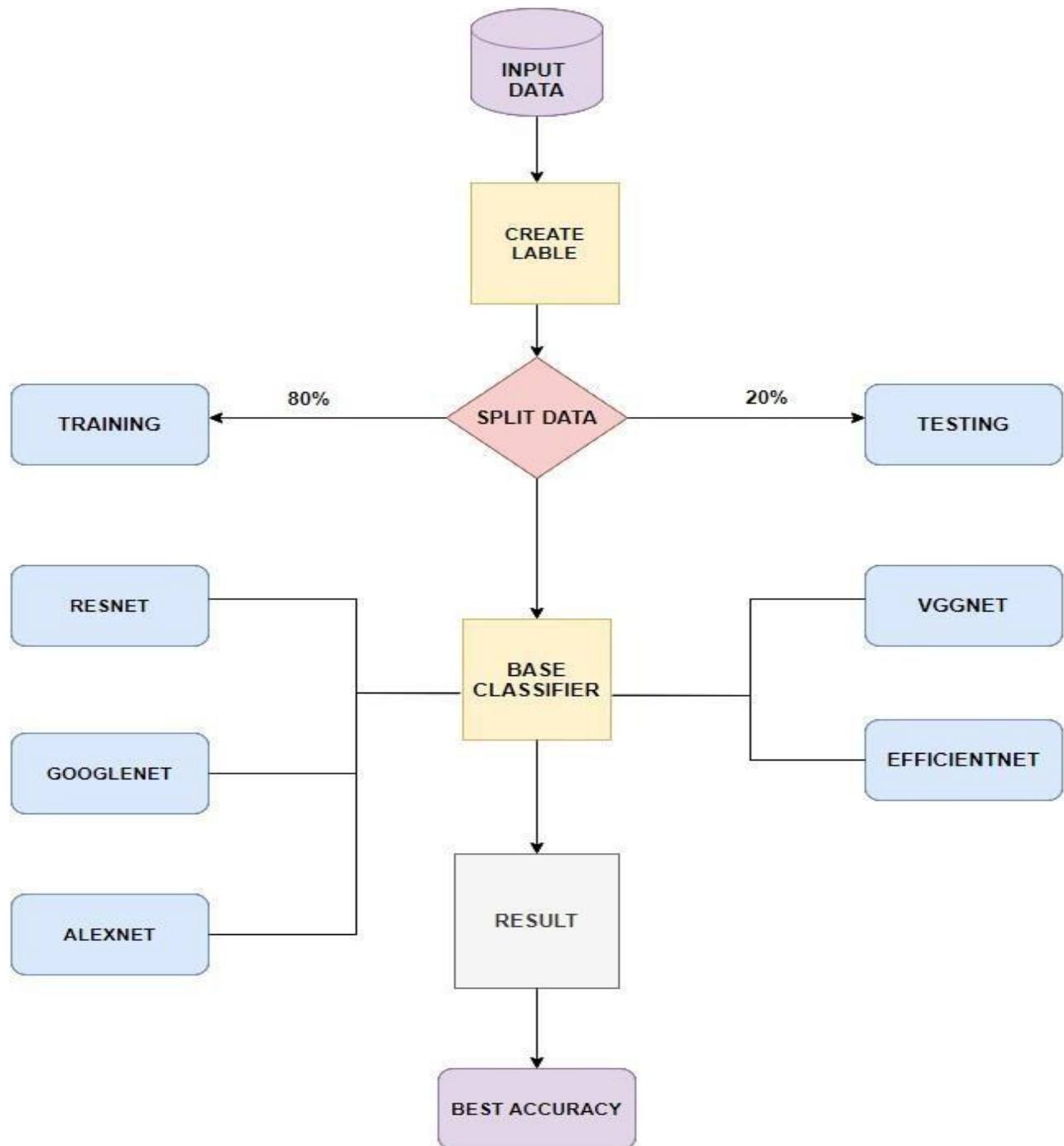
4. PROPOSED METHOD

The purpose is to identify and classify female breast cancer whether it is Malignant or Benign. This might help doctors and other medical practitioners handle patients appropriately. This could include both manageable and unmanageable risk factors in an effort to reduce, if not completely eradicate, the high death rate. . The main issue with breast cancer is how to find it. Instruments are available and ready to use. Which can predict what will happen with breast cancer in the future, but they are either expensive or poor at determining the likelihood of breast cancer. Researchers use a variety of Machine Learning algorithms and Deep Learning Models to address this classification issue. The most well-known method for classifying breast cancer are artificial neural network, which also includes random forests, XGBoost, and other techniques along with this Deep Learning models also giving most precise accuracies to classify breast cancer in this models like VGGNET, ALEXNET etc, are used. Scientists work to find the most straightforward algorithm to get the most accurate classification result, but data of varying quality will also affect the classification outcome. . Existing machine learning methods were used to develop a solution

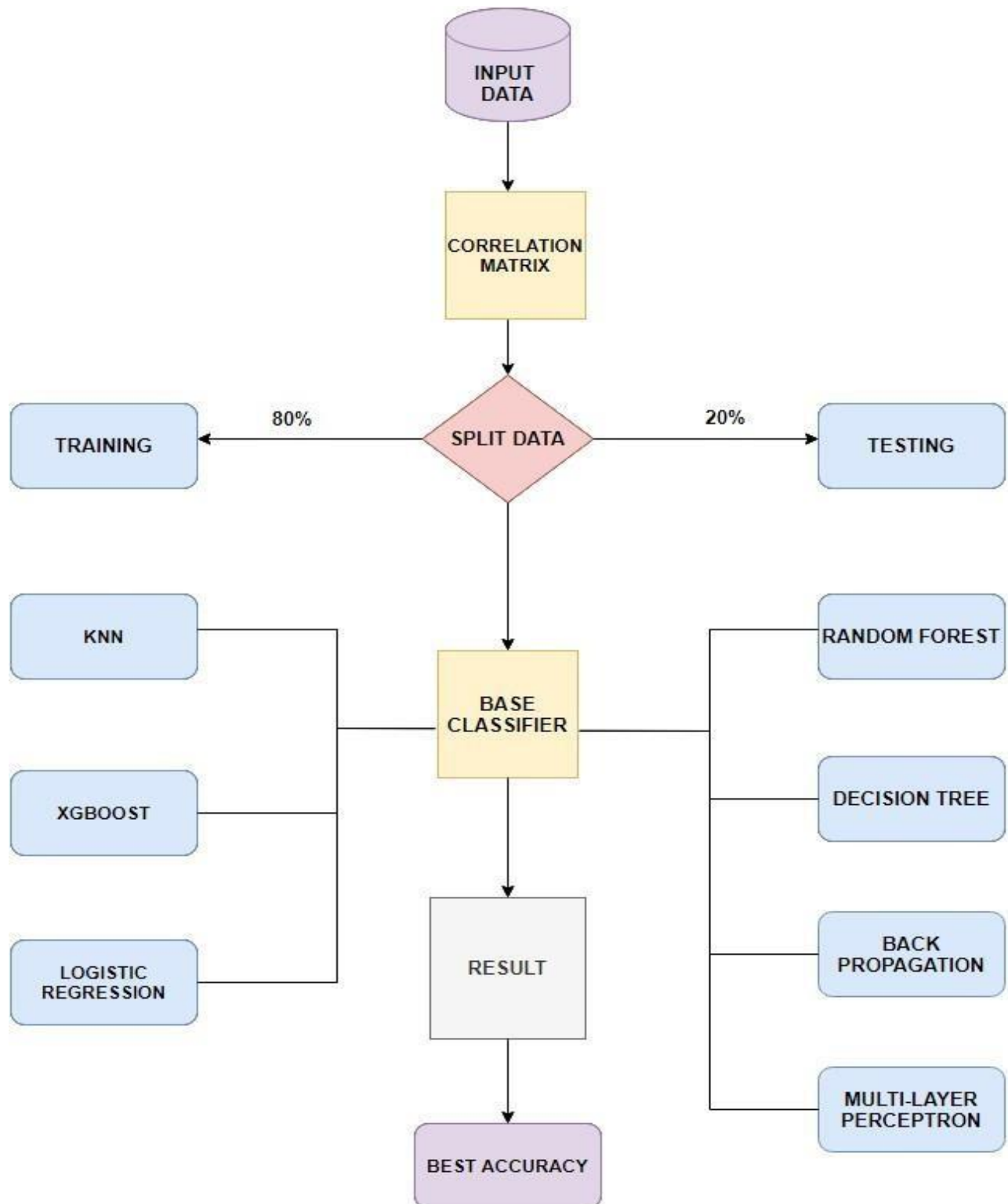
for this problem, including Random Forest, XGBoost, Decision Trees, Logistic Regression, K-Nearest Neighbor, Back Propagation, and Multi-Layer Perceptron. The data will be stored and preprocessed before being separated into train and test data, which will be fed into machine learning to learn and evaluate the model's accuracy and efficiency by classifying Breast Cancer.

4A. FLOWCHART

DEEP LEARNING FLOWCHART:



MACHINE LEARNING FLOWCHART:



5. DATA

5A DATACOLLECTON/PREPARATION:

Data preparation is necessary for any machine learning technique because the effectiveness of the methodology depends on how effectively the dataset is organised and prepared. The first five head values are shown. It eliminates redundant or noisy input and counts the number of null or duplicate values. The new data is afterwards transferred, any duplicates removed, and stored in a data frame. There are zero noisy data points. We can therefore go on to the database.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59999 entries, 0 to 59998
Data columns (total 32 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   id                                     59999 non-null  int64
1   feature_names                         59999 non-null  object
2   Radius_mean                           59999 non-null  float64
3   Texture_mean                           59999 non-null  float64
4   perimeter_mean                         59999 non-null  float64
5   area_mean                             59999 non-null  float64
6   smoothness_mean                       59999 non-null  float64
7   compactness_mean                      59999 non-null  float64
8   concavity_mean                        59999 non-null  float64
9   concave points_mean                   59999 non-null  float64
10  symmetry_mean                         59999 non-null  float64
11  fractal_dimension_mean                59999 non-null  float64
12  radius_se                             59999 non-null  float64
13  texture_se                             59999 non-null  float64
14  perimeter_se                           59999 non-null  float64
15  area_se                                59999 non-null  float64
16  smoothness_se                         59999 non-null  float64
17  compactness_se                        59999 non-null  float64
18  concavity_se                          59999 non-null  float64
19  concave points_se                     59999 non-null  float64
20  symmetry_se                           59999 non-null  float64
21  fractal_dimension_se                  59999 non-null  float64
22  radius_worst                          59999 non-null  float64
23  texture_worst                         59999 non-null  float64
24  perimeter_worst                       59999 non-null  float64
25  area_worst                            59999 non-null  float64
26  smoothness_worst                     59999 non-null  float64
27  compactness_worst                     59999 non-null  float64
28  concavity_worst                       59999 non-null  float64
29  concave points_worst                  59999 non-null  float64
30  symmetry_worst                        59999 non-null  float64
31  fractal_dimension_worst               59999 non-null  float64
dtypes: float64(30), int64(1), object(1)
memory usage: 14.6+ MB
```

We prepared a collection of 1008 images for deep learning. First, we looked at the images' structure and format to understand them better. Then, we cleaned the data by getting rid of any images that were not usable or had missing labels. To make our dataset bigger and more diverse, we used a technique called data augmentation. We split the dataset into different parts for training, validation, and testing. We also made

sure the images were ready for the model by resizing them and making the colors consistent. Finally, we made the process of loading and using the data more efficient. All of these steps were important to get our image dataset ready for deep learning.

5B DATA DESCRIPTION:

In this study, we used breast cancer dataset to create our model. We took this dataset from Kaggle. Initially there were only 5630 rows and 32 columns in the dataset taken. We extended the dataset to 60000 rows and 32 columns by adding the data from multiple resources. Our dataset is classified into two types, Benign and Malignant (cancer is present). There are 30081 Benign and 29918 Malignant. For deep learning, we took 1008 images from Kaggle.

Feature Names:

symmetry_mean: 0.002256861363739565
concavity_mean: 0.000731848460371126
area_worst: -0.00450721925846031
concave points_worst: 0.0013602646339269107
concave points_mean: 0.005354190934628159
texture_se: -0.0030119981458570067
symmetry_se: 0.0032516257958275802
perimeter_mean: 0.0012484152792571967
texture_worst: 0.001850696533796521
compactness_worst: 0.0012376152933919796
smoothness_worst: 0.0026255672285418414
concave points_se: -0.0022097991033775332
area_se: 0.006049111281941791
smoothness_mean: 0.006892435616741499
Radius_mean: -0.0019267905620734035
fractal_dimension_worst: -0.0039959011033324785
compactness_se: 0.0012903249170977372
perimeter_se: 0.0014045132646351157
Texture_mean: 0.002615765180693165
radius_se: -0.0008419542639894707
smoothness_se: 0.0015353121830169651

concavity_worst: -0.0041968767997867164
radius_worst: -0.004789520249586087
perimeter_worst: -0.004580011319336242
fractal_dimension_mean: -0.003749664503229808
symmetry_worst: 0.0008227587755407397
area_mean: 0.010072449878437924
concavity_se: 0.00290461057852096
fractal_dimension_se: -0.0013865370343537169
compactness_mean: -0.0027712526740760047

The CBIS-DDSM (Curated Breast Imaging Subset of DDSM) is an enhanced and standardized edition of the Digital Database for Screening Mammography (DDSM). The DDSM comprises 1008 film mammography studies that have been scanned and digitized. This database includes a variety of cases, including normal, benign, and malignant instances, and each case is accompanied by verified pathology information. The scale of the database along with ground truth validation makes the DDSM a useful tool in the development and testing of decision support systems.

Correlation: Correlation means, a normalized form of the covariance is a correlation between two variables. The correlation coefficients are usually between -1 and 1. The correlation coefficient is also known as Pearson's correlation coefficient. Divide the sample covariance of X and Y by the product of the sample standard deviation of X and Y, in order to get the correlation coefficient between the random variables X and Y.

$$\text{Correlation} = \frac{\text{Cov}(x,y)}{\sigma_x \cdot \sigma_y}$$

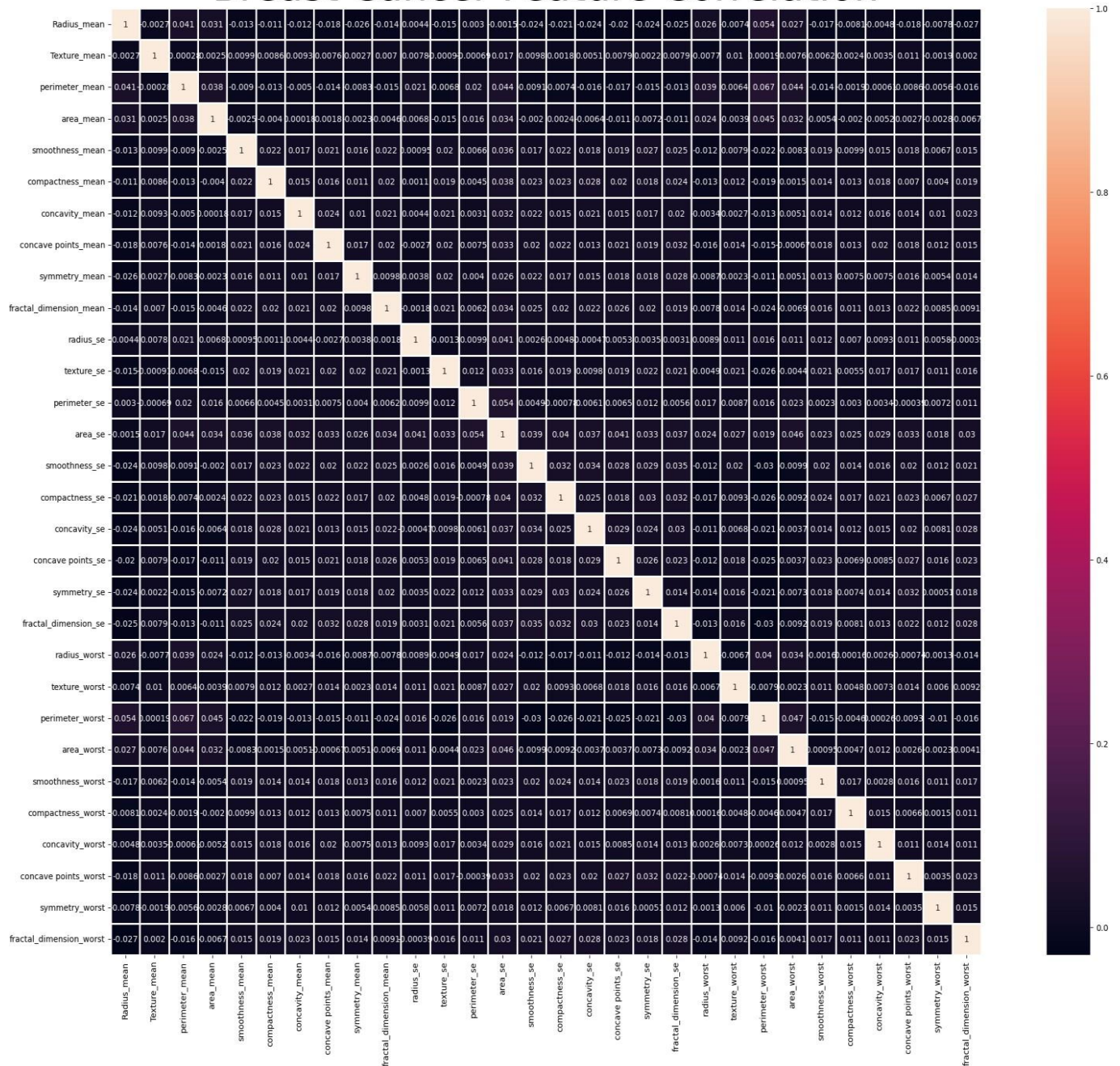
Where, Correlation = sample correlation between X and Y

Cov(X, Y) = sample covariance between X and Y

σ_x = sample standard deviation of X

σ_y = sample standard deviation of Y

Breast Cancer Feature Correlation



5C DATA ANALYSIS/VISUALIZATION:

We conducted a thorough exploratory data analysis on the aforementioned characteristics and acquired valuable information into their impact on breast cancer. We created a count plot to provide information about the data, such as if output=M, it represents a person affected by breast

cancer, and if output=B, it indicates a person who is not affected by breast cancer. Later, for calculating purposes, we substituted M for 1 and B for 0.

1 = affected by breast cancer

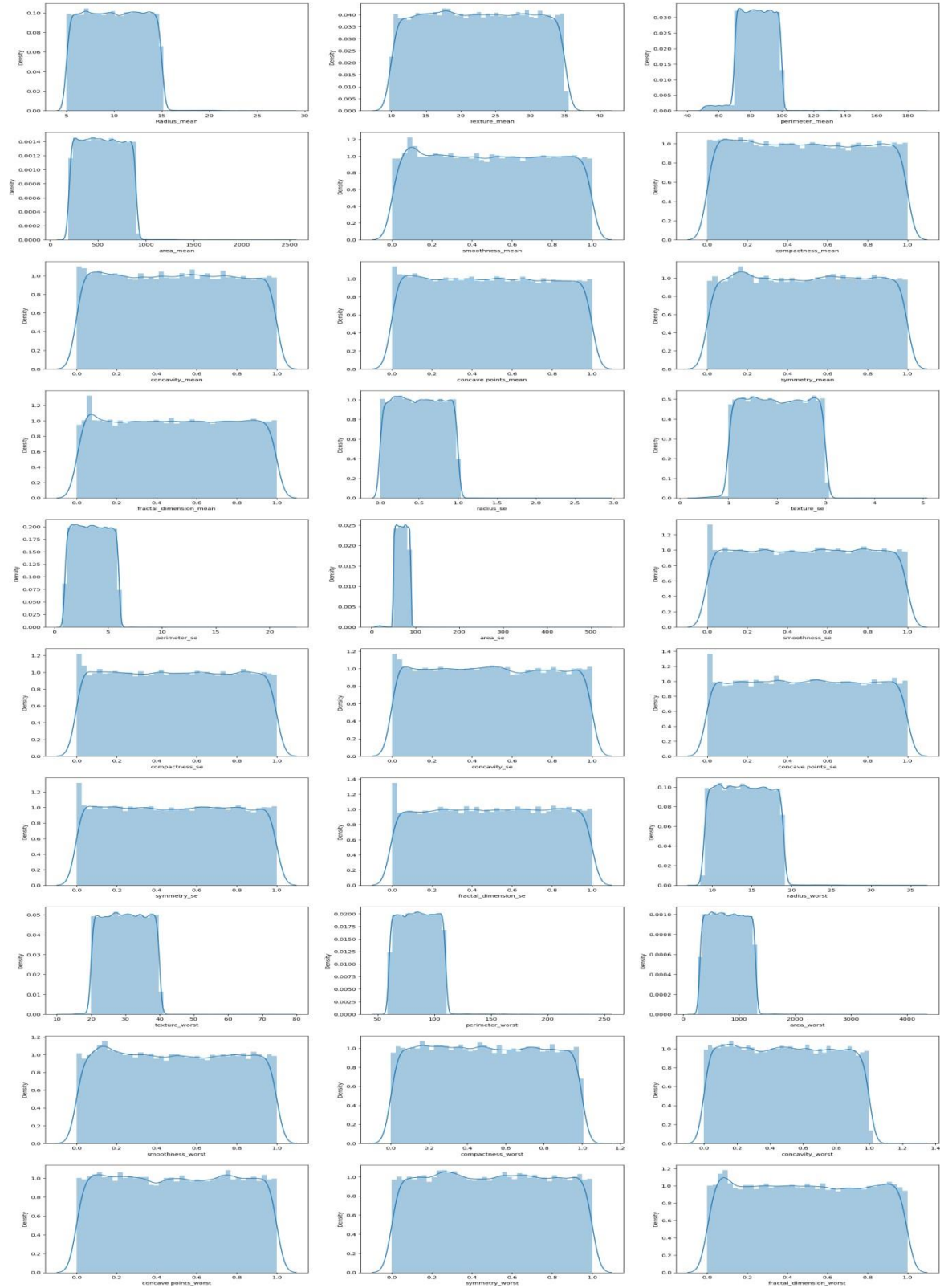
0 = not affected by breast cancer.

We then used data visualization to create histograms and box plots, making it easier to understand the dataset.

For our deep learning dataset of 1008 images, we analyzed and visualized the data. We checked if the images were classified as malignant or benign. We ensured that the number of images in each category was balanced. We also verified if any images or labels were missing. Additionally, we examined the average size of the images and compared differences between the two types. To visualize the data, we created graphs and charts showing the proportion of each type. We also displayed sample images from both categories. Through these steps, we gained a better understanding of the dataset and prepared it for deep learning.

5D VISUALIZATION PLOTS:

Histogram and Bar plots:



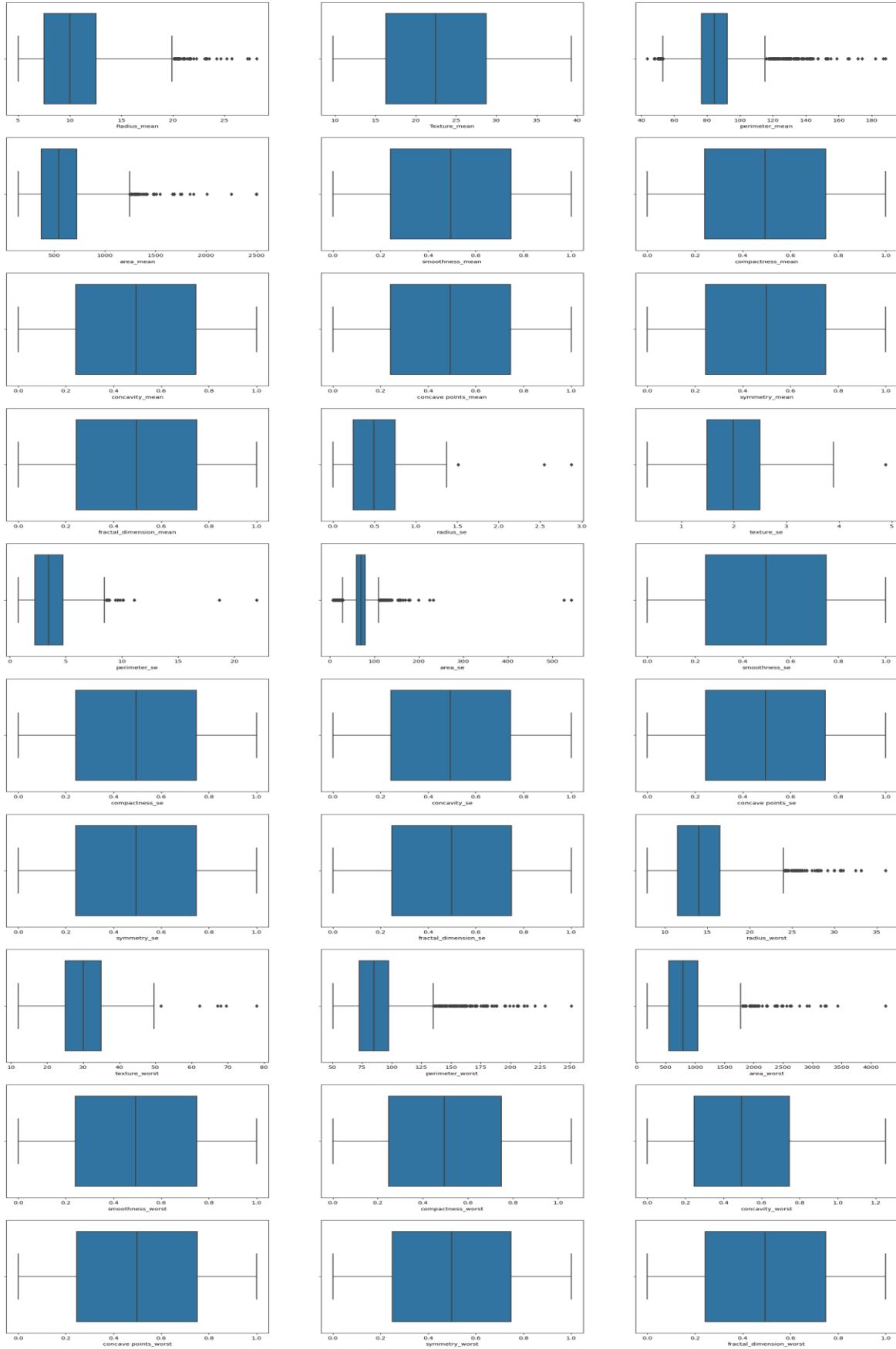
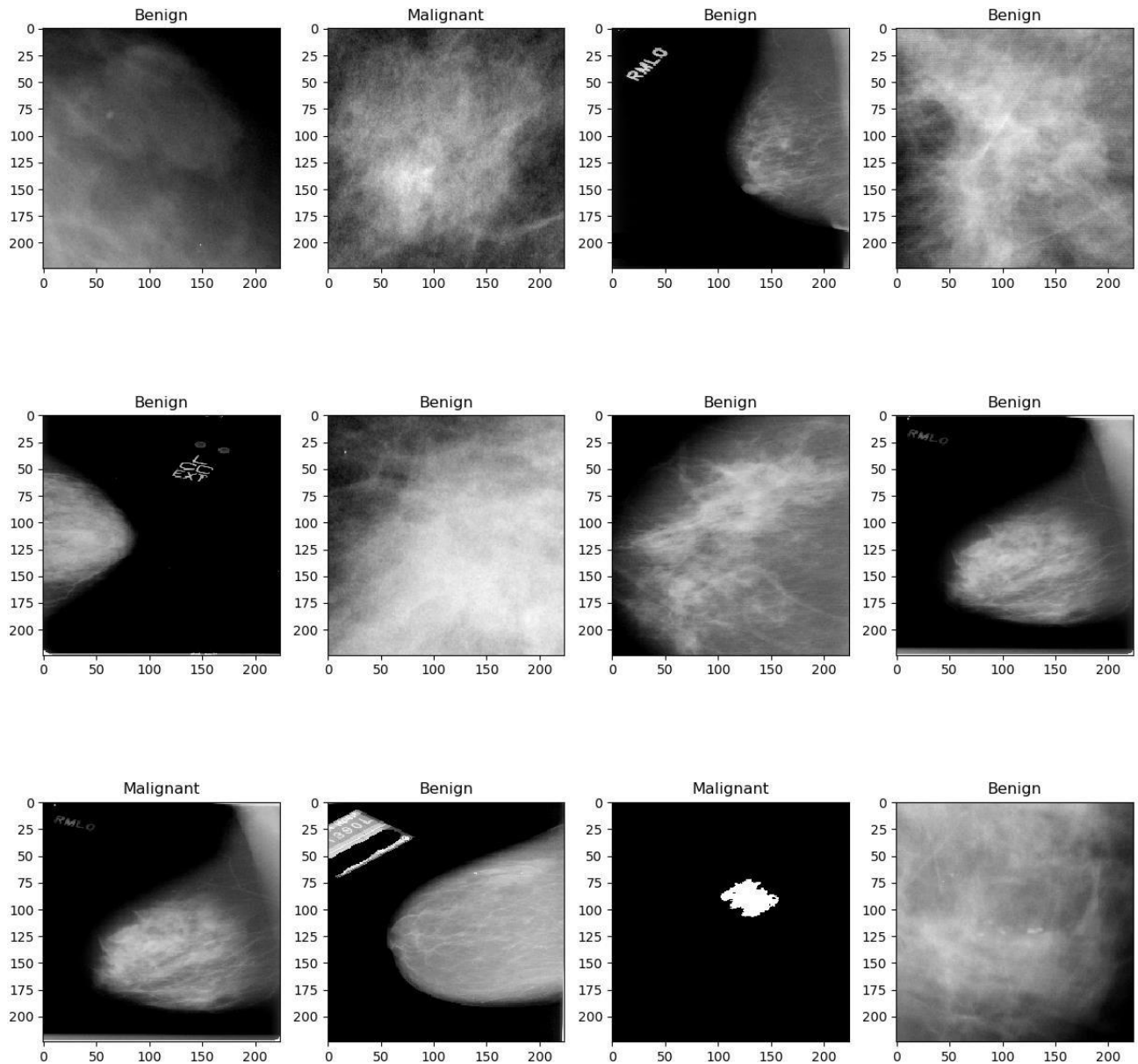


Image output visualization:



6. TESTING AND TRAINING OF MODELS

We trained various models on the dataset using well-known machine learning and deep learning classification algorithms before combining them into an ensemble model to discover the best effective approach. These categorization approaches were assessed using the following performance metrics.

Accuracy: Accuracy is the proportion of correctly predicted events from our model, or the number of correctly predicted events among all predicted events.

Confusion matrix: To assess the effectiveness of a classification model, a confusion matrix is a table that is frequently employed. By contrasting the model's predictions with the actual values in the dataset, it is especially helpful for evaluating how accurate the model's predictions are.

$$\text{Confusion Matrix} = \frac{\text{Correct Predictions}}{\text{Number of Predictions}} = \frac{TP+TN}{TP+TN+FP+FN} \text{ Where,}$$

- True Positive (TP) = Observation is positive, and is predicted to be positive.
- False Negative (FN) = Observation is positive, but is predicted negative.
- True Negative (TN) = Observation is negative, and is predicted to be negative.
- False Positive (FP) = Observation is negative, but is predicted positive

Precision: Precision in the field of information retrieval is the proportion of documents that are found that are relevant to the search.

- Precision = $\frac{TP}{TP+FP}$

Recall: A fraction of the linked records that are efficiently acquired while getting information is referred to as recall.

- Recall = $\frac{TP}{TP+FN}$

Recall is the percentage of real positives accurately categorized from all forecasts.

F- Score/F-Measure: In the statistical study of binary classification, the F-score or F-Measure is a measure of a test's accuracy.

- F-Score/F-Measure = $\frac{2*precision*recall}{precision+recall}$

In Our project the obtained accuracy during training the data was 80 % and during testing was 20%.

MODEL TRAINING AND EVALUATION USING ACCURACY SCORE

The dataset was subjected to seven (07) classification algorithms and 5 Deep Learning Models in order to determine the best performing algorithm by comparing accuracy.

The seven Algorithms are Random Forest, XGBoost, Decision Tree, Logistic Regression, K-Nearest Neighbour, Back Propagation and Multi-Layer Perceptron. Here both ML and DL algorithms were compared based on their performance evaluation metrics they are Accuracy, Precision, Recall, F-Score/F-Measure. confusion matrix was obtained to calculate the Accuracy, Precision, Recall, F-Score of the result for each algorithm. The efficiency of various methods, such as precision, recall, f-measure, and precision-recall, was compared using various statistical parameters (PRC).

MACHINE LEARNING ALGORITHMS:

Decision Tree Algorithm

Decision Tree is one of the supervised Machine Learning Algorithms and is used to solve most of the classification problems and also for prediction. It is in the form of Tree Structure. Each internal node represents a test on an attribute, each branch represents a test outcome, and each leaf node carries a class label.

$$\text{Entropy}(X) = - \sum_{i=1}^k p(a_i) \log_2 p(a_i)$$

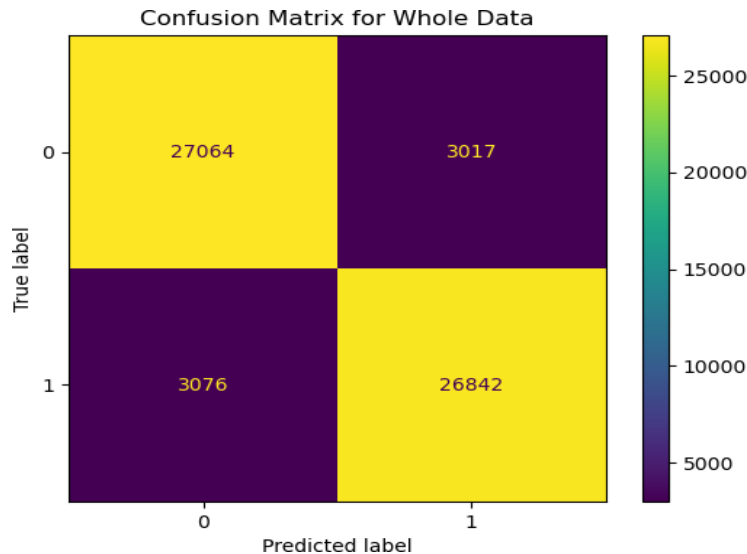
Class entropy is calculated when k is a collection of classes in a, a is the current data set to be computed where p(a) is the number of elements in class k divided by the number of elements in set a. After calculating Entropy we have to find the Information to create a tree. We take the highest information gain as a root node.

information Gain = class entropy- entropy attributes

Accuracy on whole data: 89.8448

	Precision	recall	f1-score	support
0	0.90	0.90	0.90	30081
1	0.90	0.90	0.90	29918
Accuracy			0.90	59999
Macro avg	0.90	0.90	0.90	59999
Weighted avg	0.90	0.90	0.90	59999

Representing the confusion matrix of Decision Tree algorithm:



Random Forest Algorithm

Random Forest is a Supervised Machine Learning Algorithm that could be used to handle Regression and Classification problems. Many decision trees are deployed for testing and prediction in Random Forest. Repeated testing is used to take samples from the dataset, and each sample produces its own decision tree. Random Forest incorporates elements of decision trees to make predictions. Random Forest use of a large dataset to make accurate predictions.

The breast cancer dataset was classified using a Random Forest classifier. The classifier was trained using the training set (X_train and y_train) and assessed using the whole dataset (X.values and y).

The outcomes demonstrate the way the Random Forest classifier classified the breast cancer dataset. The aggregate dataset's accuracy score indicates how accurately the predictions were made in general. The precision and recall scores reveal information on how well the classifier can detect positive samples. Taking into account both false positives and false negatives, the F1 score combines precision and recall into a single metric

Accuracy on whole data: 90.001500

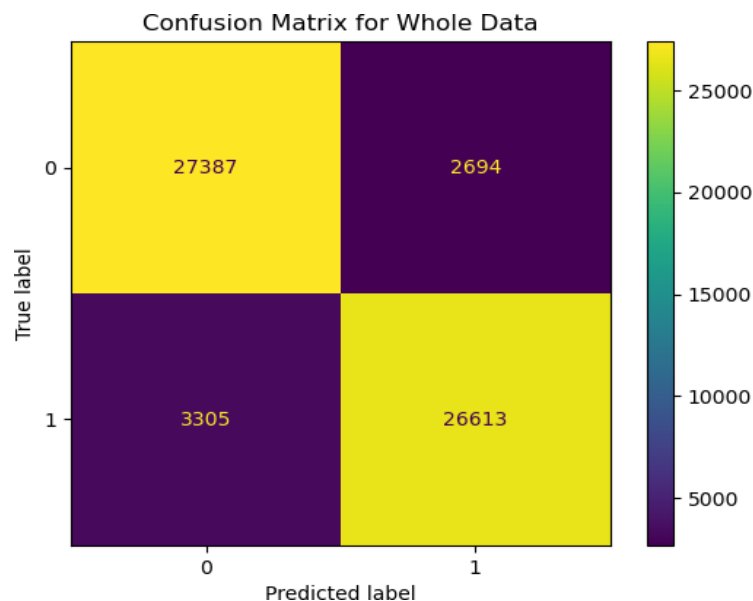
The distribution of predicted and actual class labels is visually summarized by the confusion matrix, enabling a more thorough evaluation of the model's performance. The classification report offers an in-depth analysis of the model's precision, recall, and F1-score for each class, as well as the support, which denotes the total number of samples in each class.

	Precision	recall	f1-score	support
0	0.89	0.91	0.90	30081
1	0.91	0.89	0.90	29918
Accuracy			0.90	59999
Macro avg	0.90	0.90	0.90	59999
Weighted avg	0.90	0.90	0.90	59999

Representing the confusion matrix of Random Forest algorithm:

To see the classification findings visually, a confusion matrix was created. The predictions generated by the Random Forest classifier are broken down as follows:

Confidence Matrix for the Whole Data:



XGBoost Algorithm

XGBoost is a decision tree based ensemble Machine Learning algorithm that uses a gradient boosting framework. It can be used to solve regression, classification, ranking and user-defined prediction problems. Popular boosting algorithms include gradient boosting. In gradient boosting, each prediction fixes the mistake made by its previous. Instead than adjusting the weights of the training instances like Adaboost does, each predictor is trained using the labels from the residual errors of the predecessor.

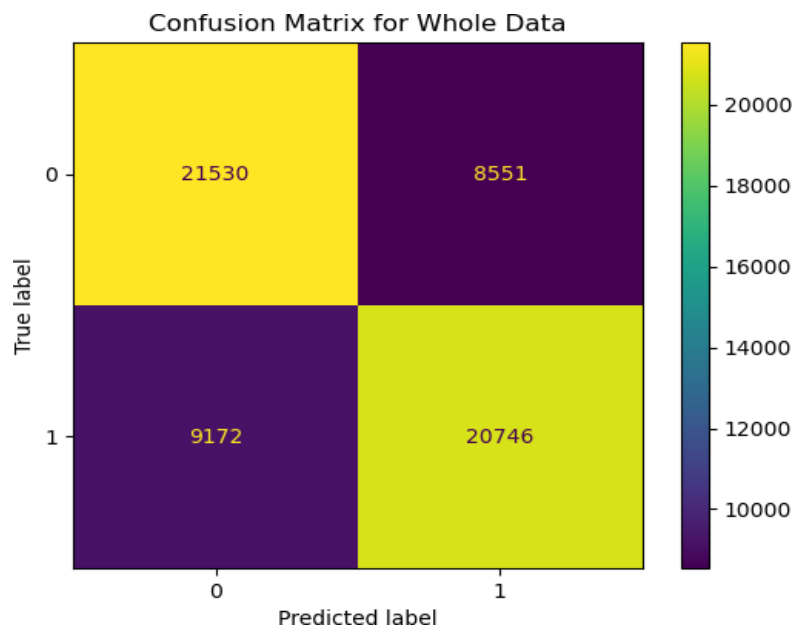
Accuracy on whole data: 70.461174

	Precision	recall	f1-score	support
0	0.70	0.72	0.71	30081
1	0.71	0.69	0.70	29918
Accuracy			0.70	59999
Macro avg	0.70	0.70	0.70	59999
Weighted avg	0.70	0.70	0.70	59999

Representing the confusion matrix of XGBOOST algorithm:

A confusion matrix was generated to visualize the classification results:

Confusion Matrix for Whole Data:



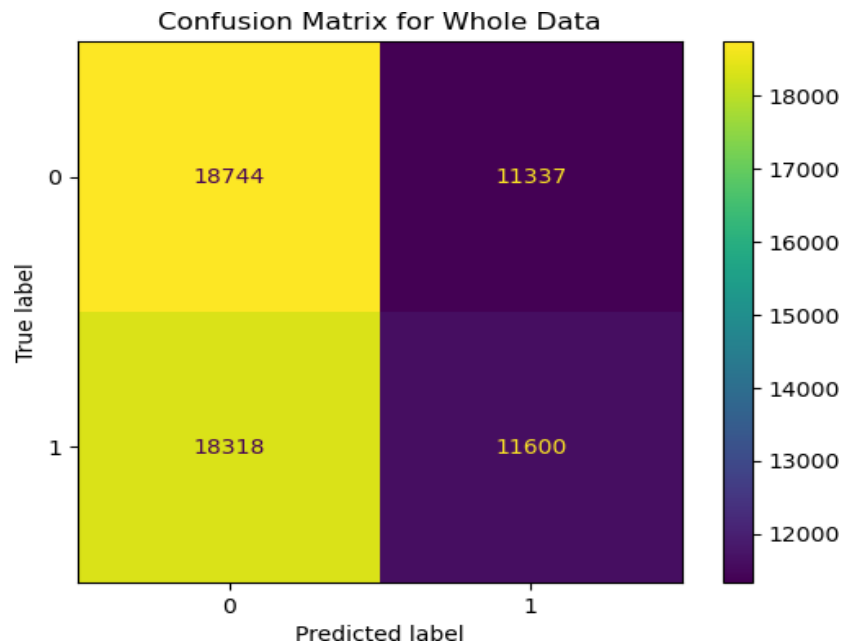
Logistic Regression Algorithm

Logistic Regression is a Supervised Machine Learning Algorithm for predicting the probability of a relationship between dependent and independent variables. The basic goal is to find the best-fitting model for relationships between independent and dependent variables. Dependent Variable is Binary means the output can be M (Malignant) or B (Benign) and Independent variable can be both continuous and binary.

Accuracy on Whole data: 50.57417

	Precision	recall	f1-score	support
0	0.51	0.62	0.56	30081
1	0.51	0.39	0.44	29918
Accuracy			0.51	59999
Macro avg	0.51	0.51	0.50	59999
Weighted avg	0.51	0.51	0.50	59999

Representing the confusion matrix of Logistic Regression algorithm:



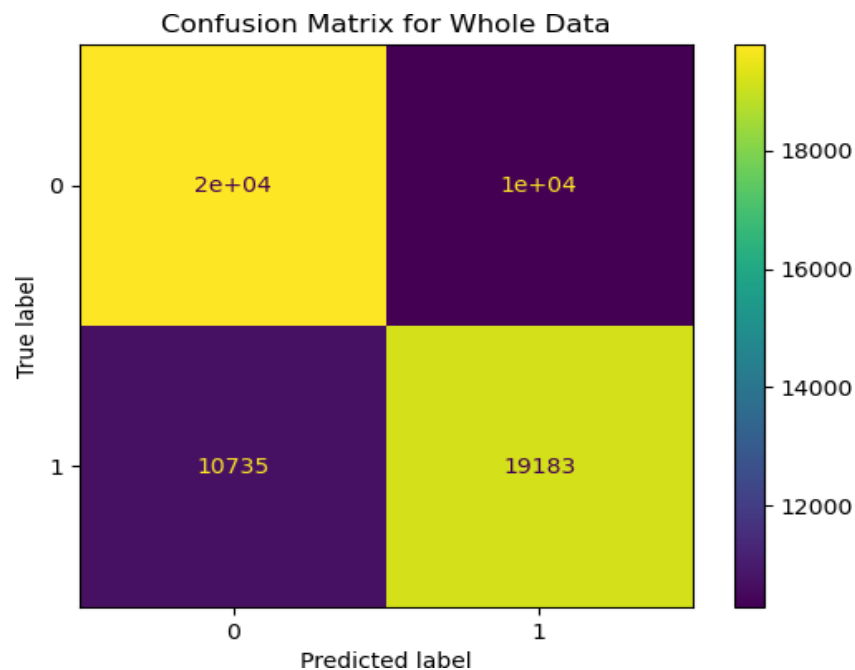
K-Nearest Neighbor (KNN)

K-Nearest Neighbor is one of the simplest Supervised Machine Learning algorithm mostly used for classification. It classifies a data point based on how its neighbors are classified. KNN stores all available cases and classifies new cases based on a similarity measure. K in KNN is a parameter that refers to the number of nearest neighbors to include in the majority voting process.

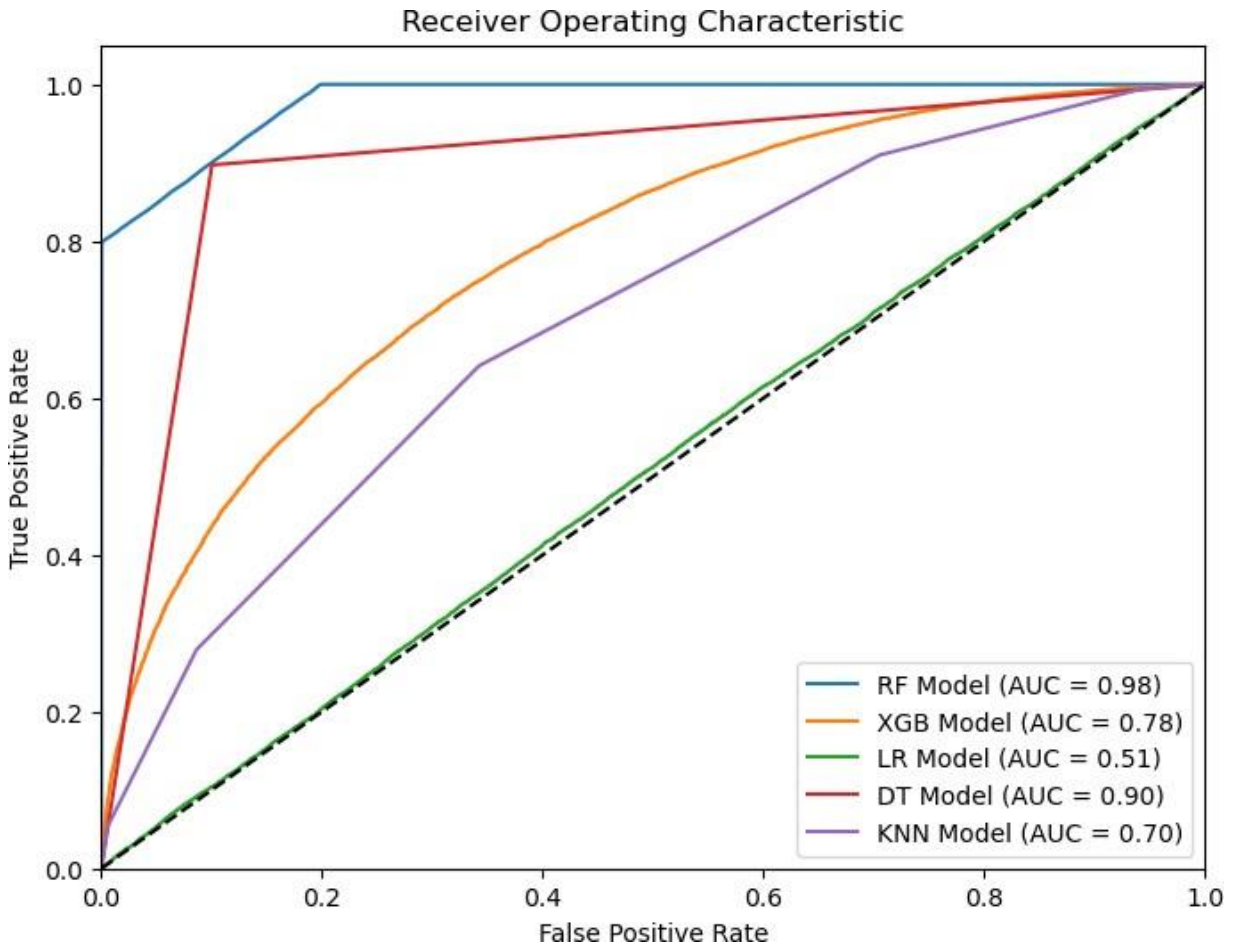
Accuracy on whole data: 64.93941

	Precision	recall	f1-score	support
0	0.65	0.66	0.65	30081
1	0.65	0.64	0.65	29918
Accuracy			0.65	59999
Macro avg	0.65	0.65	0.65	59999
Weighted avg	0.65	0.65	0.65	59999

Representing the Confusion Matrix of K-Nearest Neighbor:



AUC AND ROC curve for all 5 algorithms:



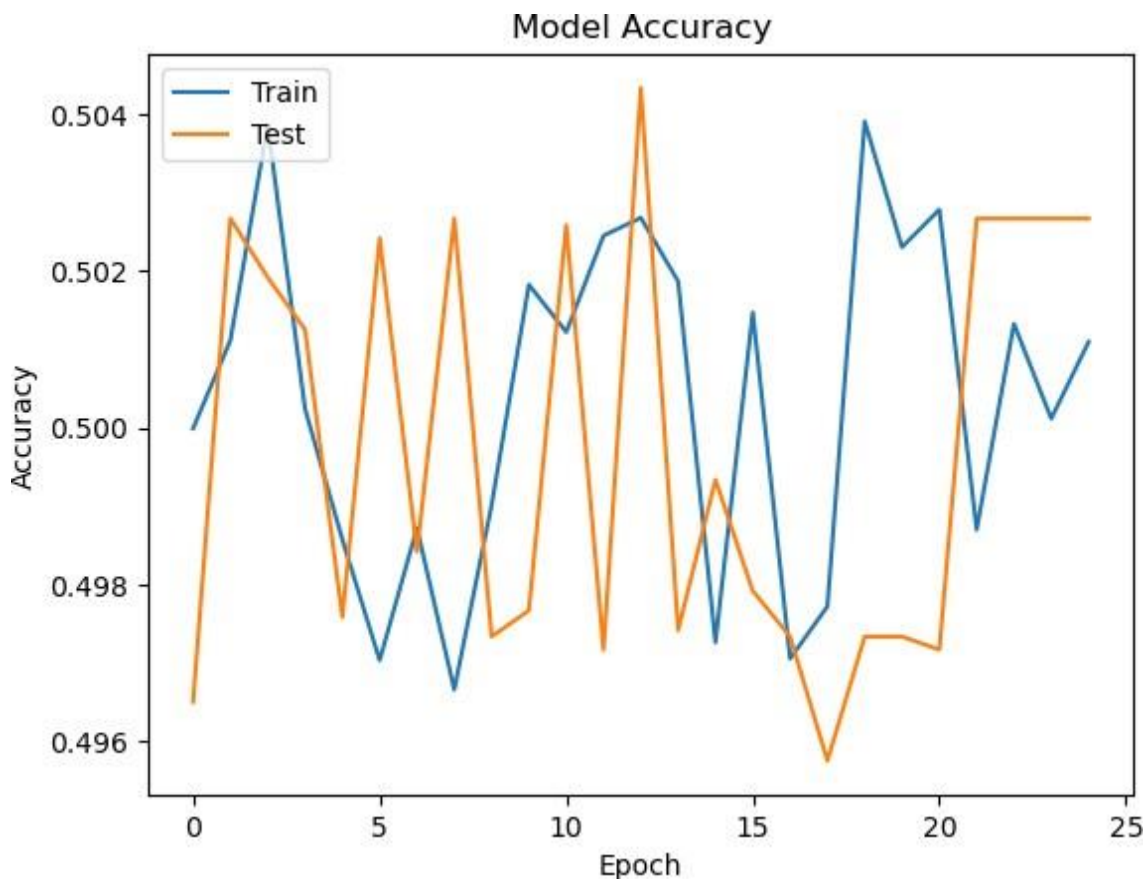
BACK PROPAGATION

An sequential model with two dense layers makes up this model architecture. For binary classification, the first layer contains 32 units with ReLU activation functions while the second layer has a single unit with a sigmoid activation function.

The Adam optimizer and the binary cross-entropy loss function are used to build the model. Using the training data, the model is trained over a period of 25 epochs with a batch size of 32. To assess the model's effectiveness during training, the validation data (X_{test} and y_{test}) are employed. After training, the model's effectiveness on the test data is assessed, and the accuracy on the test data is printed. This accuracy rating shows how well the model can identify breast cancer instances based on hidden data.

To further illustrate the model's accuracy over the epochs for both the training and validation sets, the learning curve is presented. The plot sheds light on the model's training development, including any over- or underfitting propensities.

Overall, this neural network model employs the Back Propagation technique to discover the most suitable weights and biases for precisely classifying breast cancer. The goal of the training process is to increase the model's precision while limiting the loss function. Understanding the model's performance and training dynamics is made possible by the evaluation on the test data and the learning curve graphic.



MULTI-LAYER PERCEPTRON

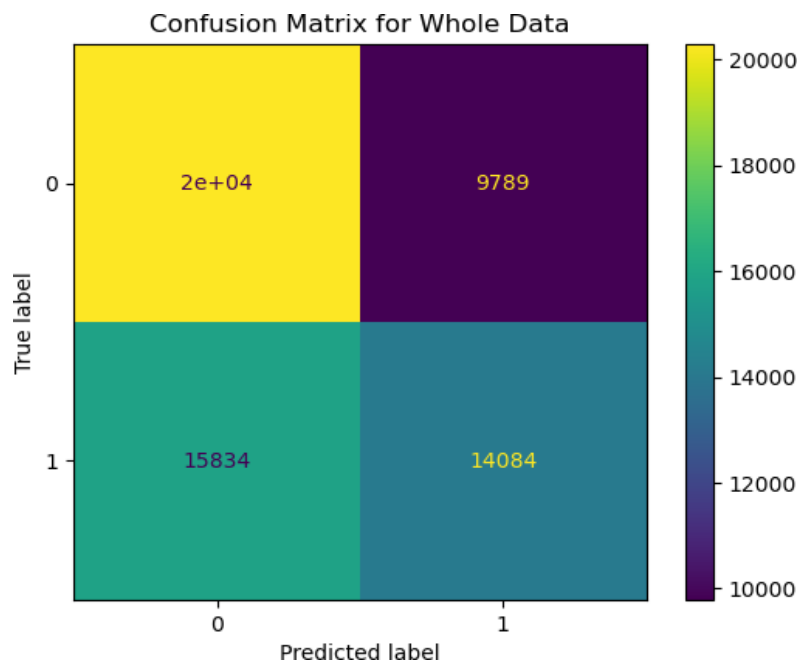
Multi-Layer Perceptron is an input layer, a hidden layer, and an output layer make up a straightforward neural network. There are various hidden layers used in deep learning. For accuracy and precisely detecting the layers in the image, several hidden layers are important and reliable. GPU is a more efficient computing platform than

CPU for these operations. The gradient decline is so minimal because of the many levels that it vanishes. The gradients start to disappear and becoming very modest in comparison to the weights of the networks when the information is transmitted back.

Accuracy on whole data: 57.830963

	Precision	recall	f1-score	support
0	0.56	0.67	0.61	30081
1	0.59	0.47	0.52	29918
Accuracy			0.57	59999
Macro avg	0.58	0.57	0.57	59999
Weighted avg	0.58	0.57	0.57	59999

Representing the confusion matrix of Multi-Layer Perceptron:



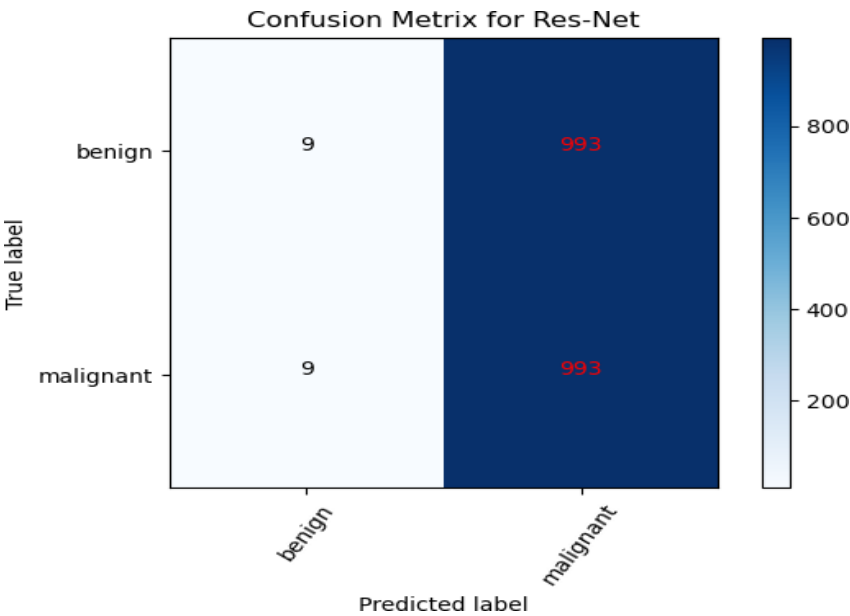
DEEP LEARNING MODELS

RESNET

ResNet, also known as a deep learning architecture called a residual neural network, was created expressly to overcome the difficulties associated with training very deep

neural networks. In this project, the ResNet model is employed as one of the deep learning models to classify breast cancer. By contrasting the actual labels (Y_test) with the predicted labels (Y_pred_tta), the confusion matrix is used to gauge how well the ResNet model is performing.

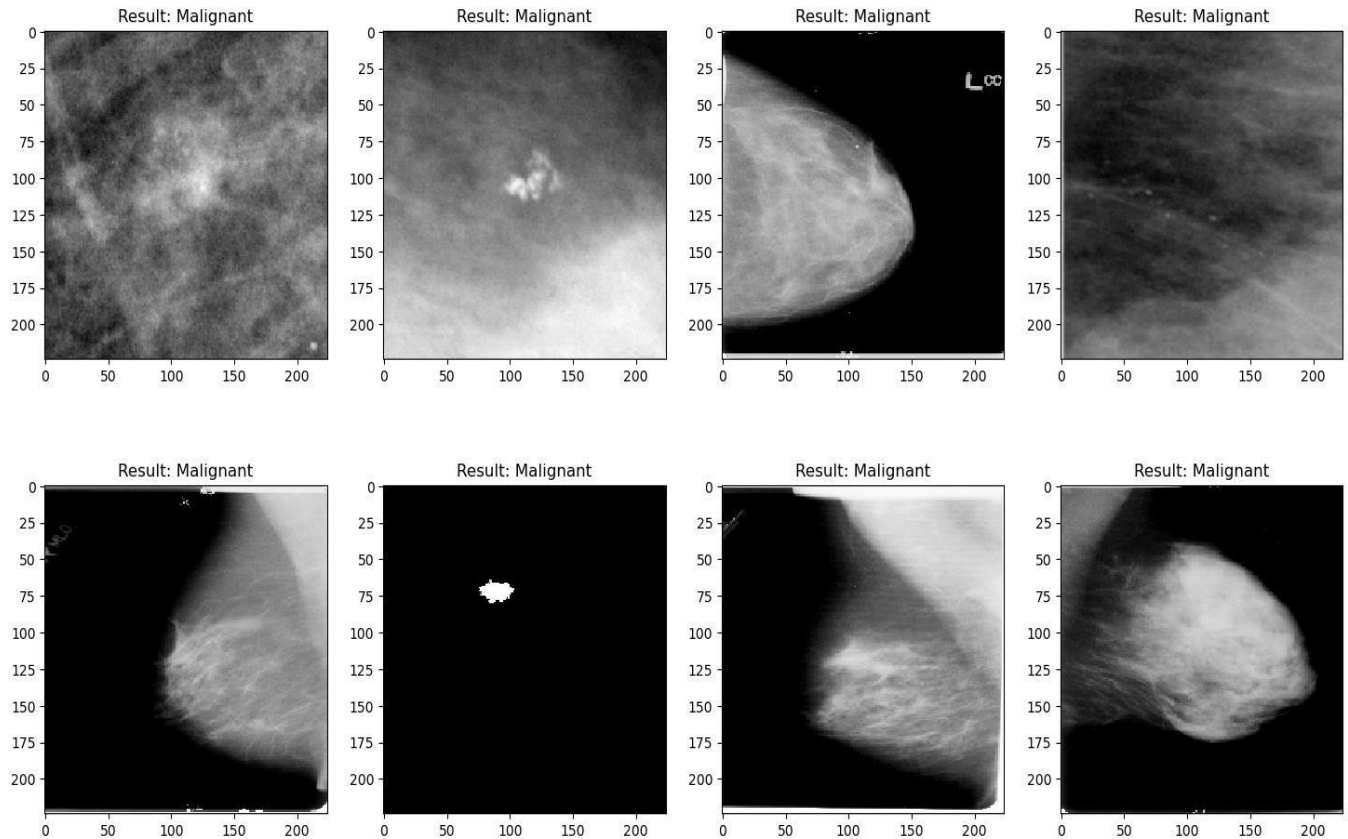
The confusion matrix provides a tabular depiction of the model's performance, displaying the numbers of accurate predictions—both positive and negative—as well as incorrect positive and negative ones. The plot_confusion_matrix function is used to visualize the confusion matrix. It takes the confusion matrix (cm), the class labels (cm_plot_label), and the title as input parameters. With the true labels on the y-axis and the predicted labels on the x-axis, the function renders the confusion matrix as a heatmap. The heatmap's colour intensity corresponds to how many predictions fell into each category.



	Precision	recall	f1-score	support
0	0.49	0.02	0.03	1002
1	0.50	0.98	0.66	1002
Accuracy			0.50	2004
Macro avg	0.49	0.50	0.35	2004
Weighted avg	0.49	0.50	0.35	2004

Accuracy on whole data: 0.528678

This model has a part where the first eight correctly identified benign photos (prop_class) and incorrectly identified benign images (mis_class) are displayed. These visualisations offer information about the model's performance and can be used to spot any patterns or difficulties that may have arisen throughout the classification process.



The confusion matrix and image visualisations aid in assessing the effectiveness of the model and providing understanding of its predictions. The model can learn and represent complicated patterns in the data thanks to the ResNet architecture, which is renowned for its skip connections and residual blocks. This improves the classification accuracy of breast cancer.

ALEXNET

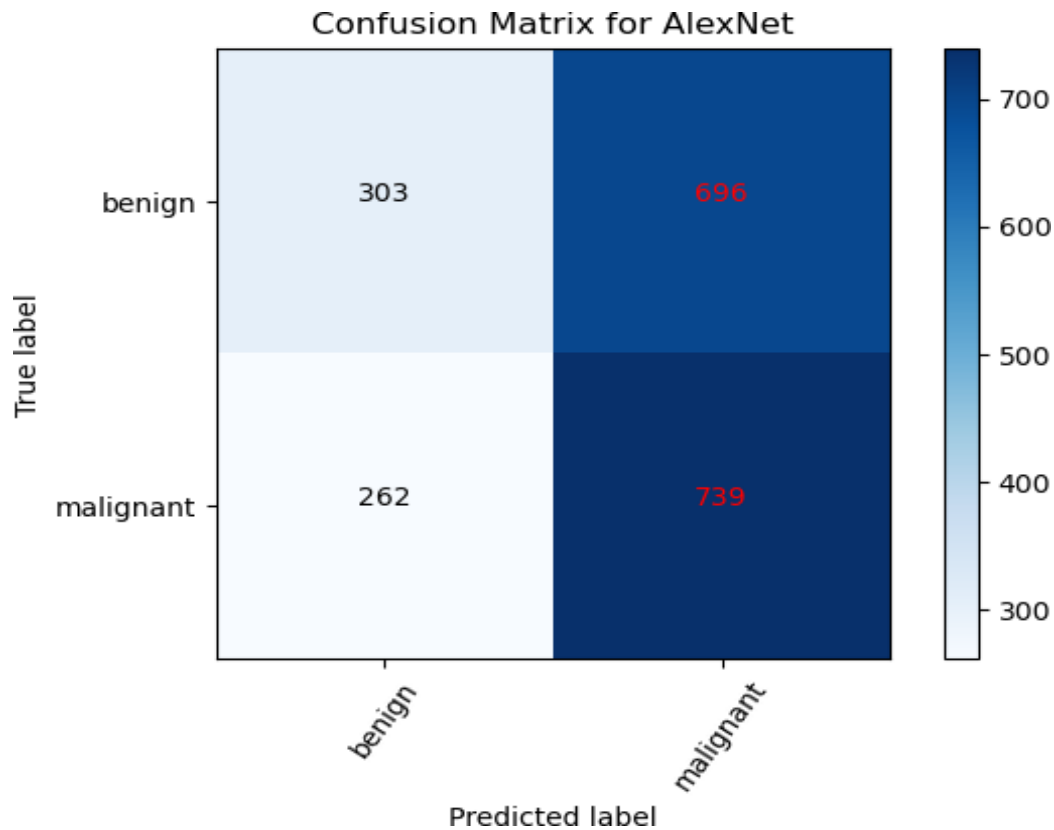
Deep convolutional neural network AlexNet made important advancements in computer vision tasks, particularly picture categorization. Utilising TensorFlow Keras' Sequential API, the AlexNet model is built in this project, It has eight layers, five of which are convolutional and three of which are fully linked. The architecture is summarised as follows:

1. Convolutional Layer 1: 96 filters with a kernel size of 11x11, stride of 4x4, and ReLU activation. This layer extracts low-level features from the input images.
2. MaxPooling Layer 1: Max pooling with a pool size of 3x3 and stride of 2x2 to downsample the feature maps.
3. Convolutional Layer 2: 256 filters with a kernel size of 5x5 and ReLU activation. This layer further captures more complex features.
4. MaxPooling Layer 2: Max pooling with a pool size of 3x3 and stride of 2x2.
5. Convolutional Layers 3, 4, and 5: These layers consist of 384, 384, and 256 filters, respectively, with a kernel size of 3x3 and ReLU activation. They continue to extract more abstract features.
6. MaxPooling Layer 3: Max pooling with a pool size of 3x3 and stride of 2x2.
7. Flatten: Flattens the feature maps into a 1D vector to connect to the fully connected layers.
8. Fully Connected Layers 6 and 7: Each layer has 4096 units with ReLU activation. Dropout is applied with a rate of 0.5 to prevent overfitting.
9. Output Layer: A dense layer with the number of units equal to the number of classes (2 in this case) and a softmax activation function for multi-class classification.

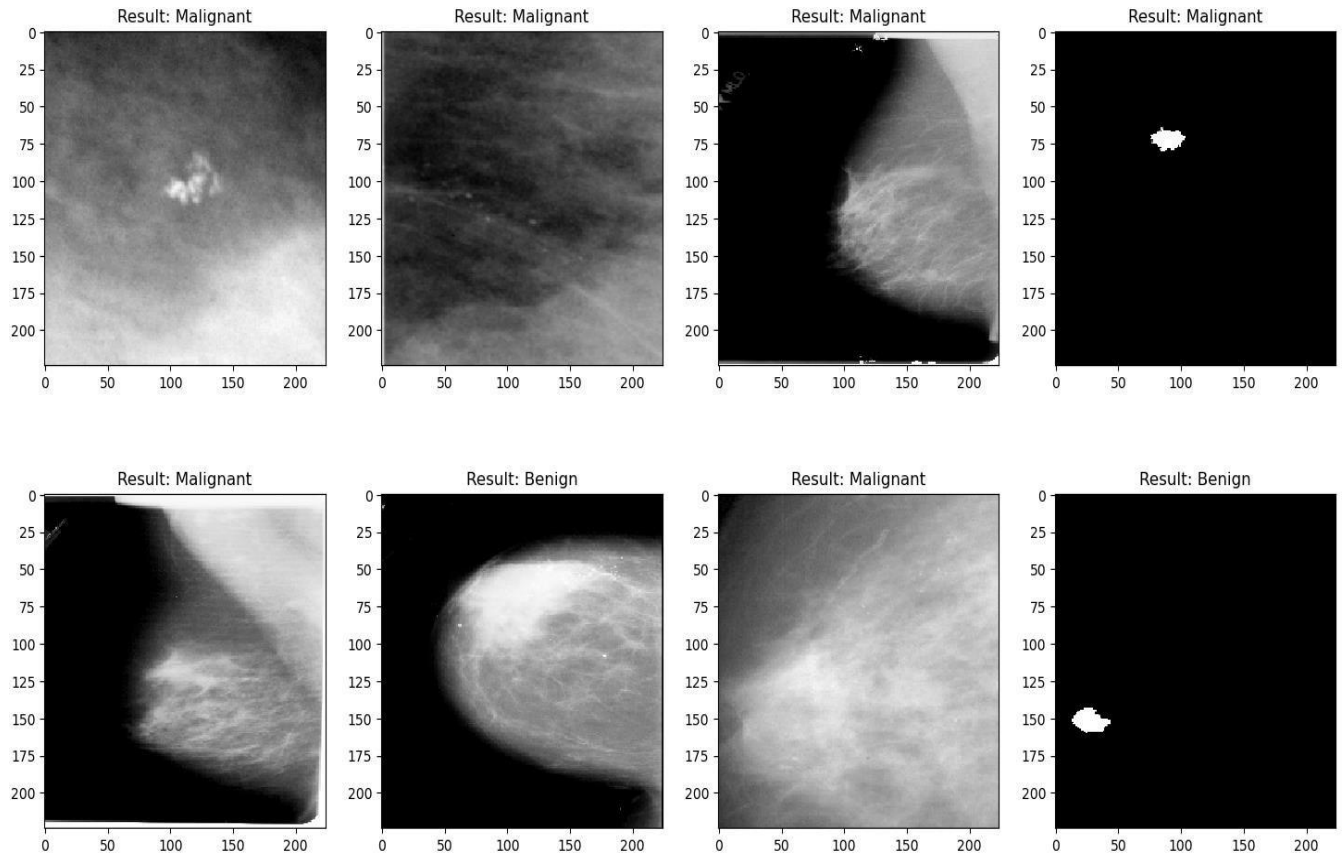
The performance of the AlexNet model in classifying breast cancer as either benign or malignant. The confusion matrix provides a visual summary of the model's predictions, allowing for a quick assessment of its accuracy and potential misclassifications. Additionally, the normalization option provides insights into the distribution of predictions across the classes.

Accuracy on whole data: 0.4937655

	Precision	recall	f1-score	support	
	0	0.54	0.30	0.39	999
	1	0.51	0.74	0.61	1001
Accuracy				0.52	2000
Macro avg		0.53	0.52	0.50	2000
Weighted avg		0.53	0.52	0.50	2000

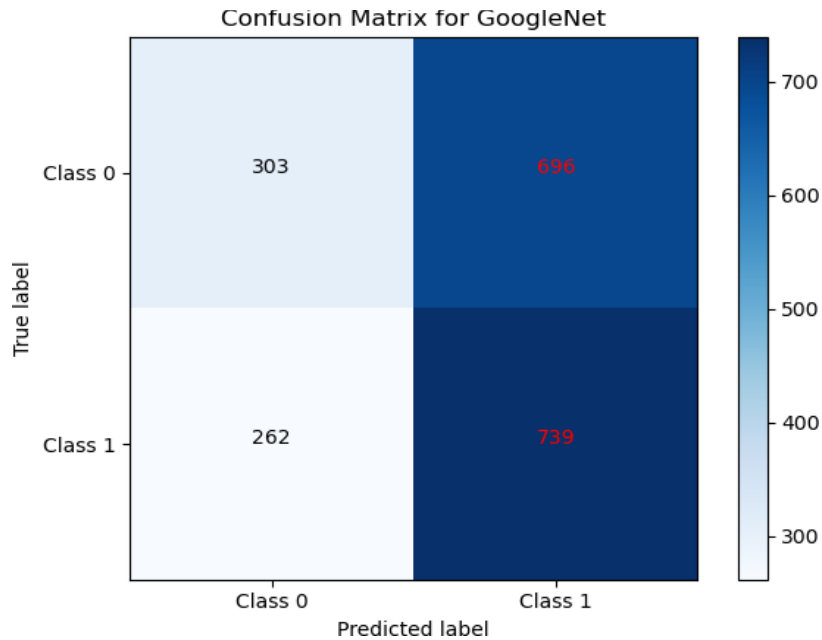


The resulting plot shows the first 8 photos of benign breast cancer patients in a visual manner, emphasizing how well the algorithm predicted these particular samples. The display gives insights into the model's performance and illustrates how well it performs in correctly identifying benign cases.



GOOGLNET

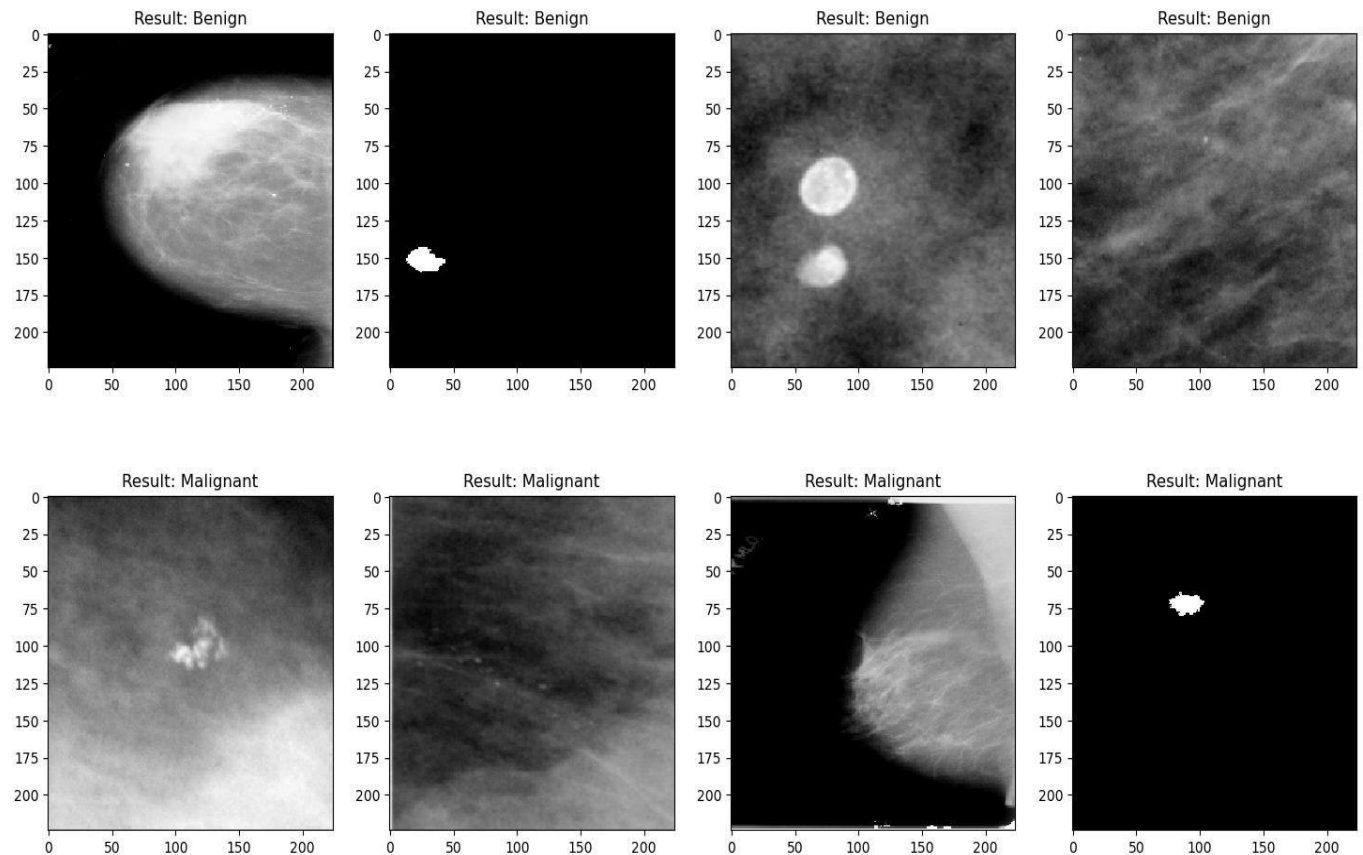
Deep convolutional neural network architecture GoogleNet is well-known for its efficiency in classification of images tasks. GoogleNet makes use of a special module known as the Inception module, which combines filters of various sizes to collect data at multiple scales. The GoogleNet model is built using the Inception modules by the `build_googlenet_model` method in this project. It returns the assembled model after receiving the input shape and number of classes as parameters. Multiple convolutional layers, max pooling layers, and Inception modules with various filter sizes make up the model. An average pooling layer, a dropout layer for regularization, and a fully connected layer with softmax activation for classification are used as the final layers. The function calculates the confusion matrix between the real labels (`Y_test_subset`) and the predicted labels (`Y_pred_tta_subset`) in order to provide a confusion matrix particular to GoogleNet's performance.



	Precision	recall	f1-score	support
0	0.54	0.30	0.39	999
1	0.51	0.74	0.61	1001
Accuracy			0.52	2000
Macro avg	0.53	0.52	0.50	2000
Weighted avg	0.53	0.52	0.50	2000

Accuracy on whole data: 0.468178

GoogleNet concentrates on showing a portion of accurately identified photos. It recognizes and records the indexes of accurately predicted benign and malignant samples in distinct lists. The success of classifying breast cancer photos is then demonstrated by using subplots to visually represent the first four photographs of each class. Additionally, the cases of benign and malignant breast cancer are visually represented in the exhibited photographs, supporting the model's capacity to distinguish between the two classes appropriately.



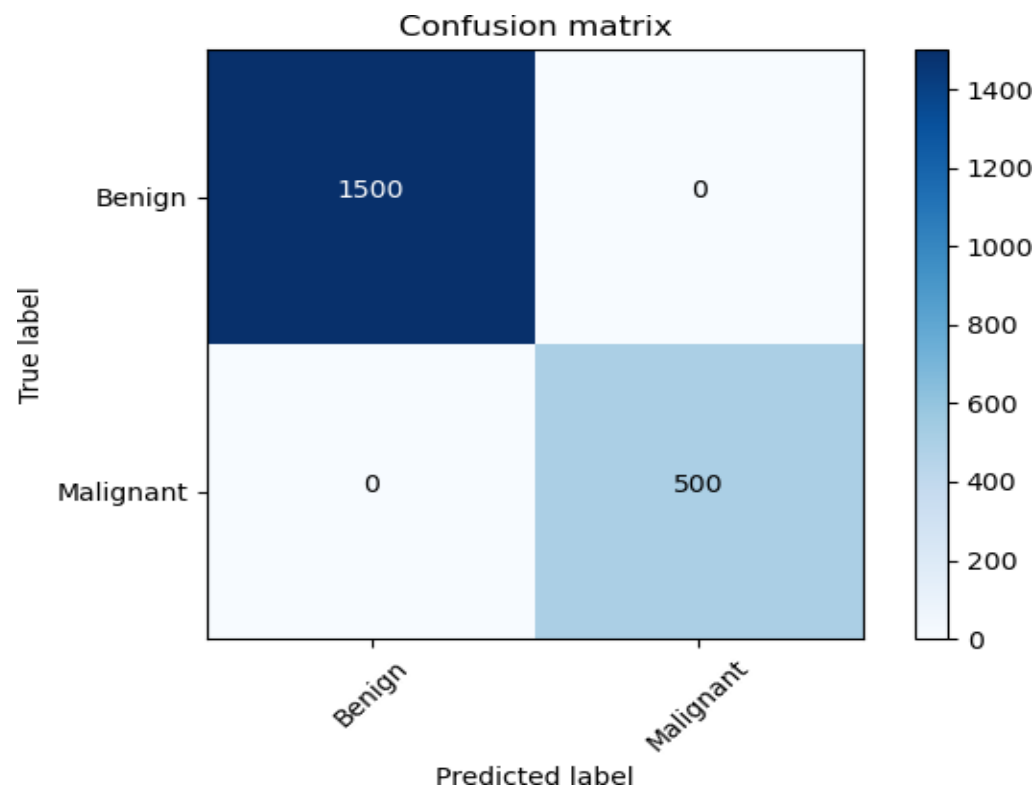
The architecture of GoogleNet, the Inception modules, and the visualization of findings show the potential of deep learning models to support medical professionals' decision-making and diagnosis of breast cancer.

VGGNET

For image classification tasks, a well-liked convolutional neural network design called VGGNet is frequently employed. The deep architecture of VGGNet, which includes stacked convolutional layers, is well renowned for allowing it to extract intricate patterns and information from images. The VGGNet model is built using the `build_vggnet_model` function. The pre-trained VGG16 model is loaded from the Keras applications module, its weights are frozen, and additional layers are added on top for fine-tuning. Convolutional layers are followed by completely connected layers

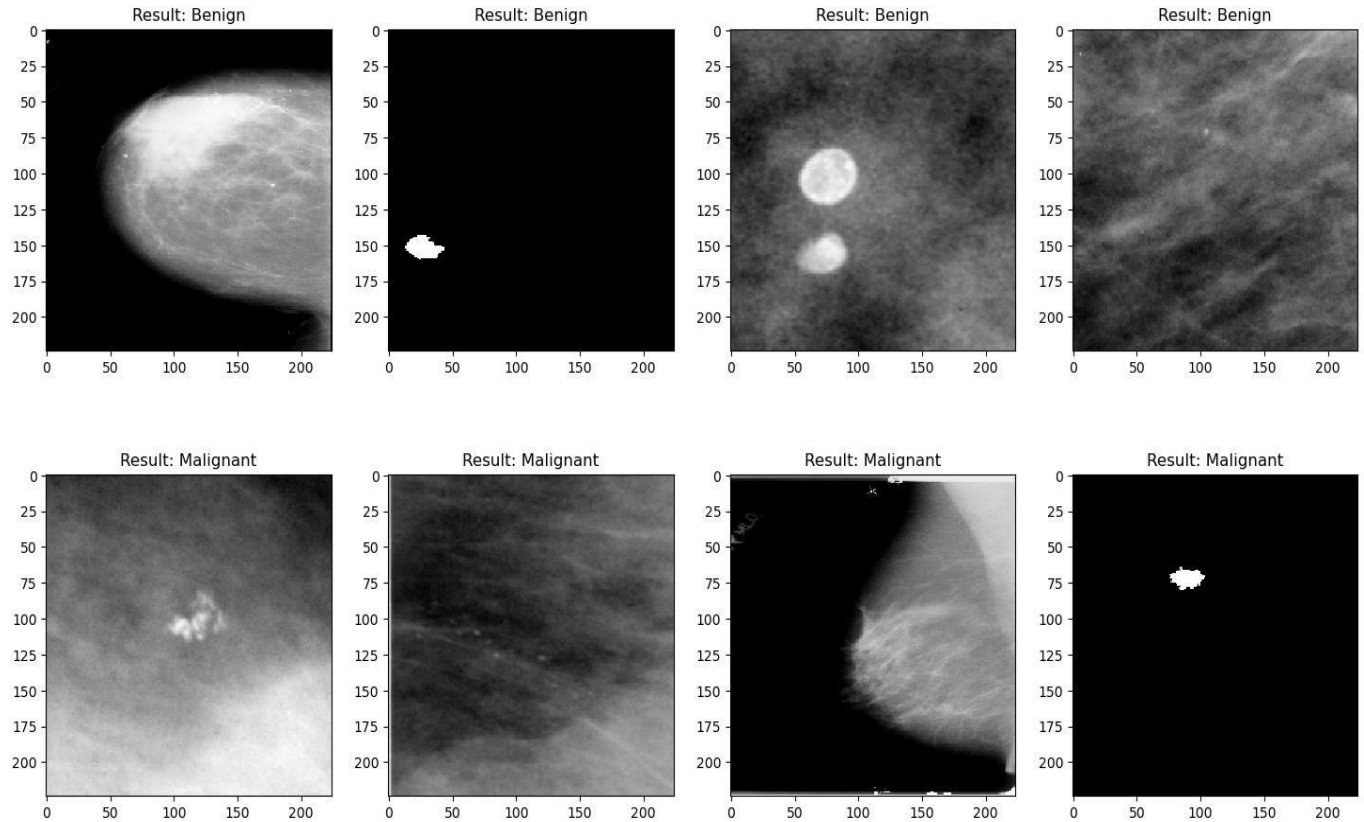
with dropout for regularization, fully connected layers with softmax activation for classification, and finally a dense layer with softmax activation.

The efficiency of VGGNet in categorizing images of breast cancer. The confusion matrix, which displays the true positive and true negative rates, sheds light on the model's effectiveness. The capacity of the algorithm to reliably distinguish between benign and malignant breast cancer cases is further supported by the visualization of correctly categorized images.



	Precision	recall	f1-score	support	
	0	0.67	1.00	0.80	2
	1	0.00	0.00	0.00	1
	2	0.67	0.67	0.67	3
Accuracy				0.67	6
Macro avg		0.44	0.56	0.49	6
Weighted avg		0.56	0.67	0.60	6

Accuracy on whole data: 0.835162



VGGNet is an effective tool for medical experts in the diagnosis of breast cancer because of its complex architecture and fine-tuning capability. VGGNet shows the promise of deep learning methods in assisting precise and effective classification jobs in the area of medical imaging by utilizing pre-trained models and transfer learning techniques.

EFFICIENT NET

This project explains to create an image classification model using the EfficientNetB0 architecture. Known for its effectiveness and outstanding performance, the EfficientNetB0 model was pre-trained using the ImageNet dataset.

The model architecture consists of the following layers:

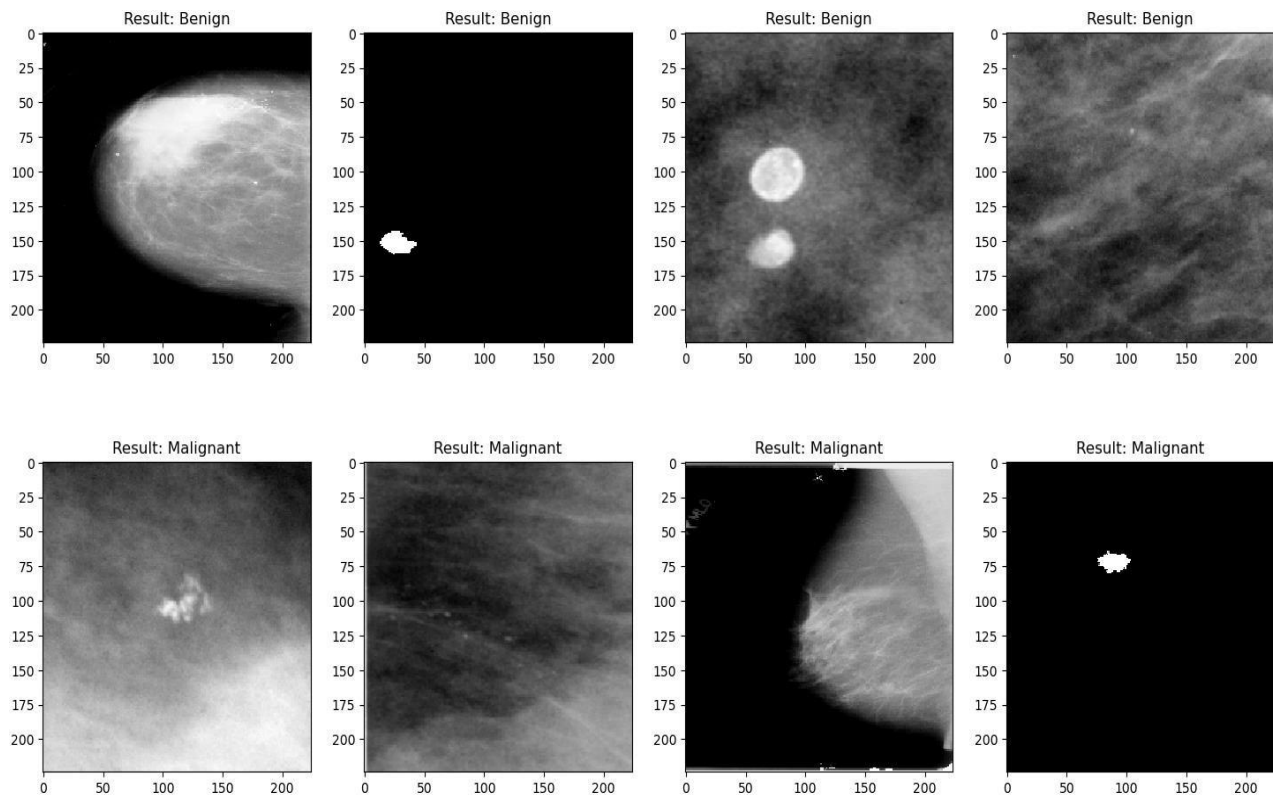
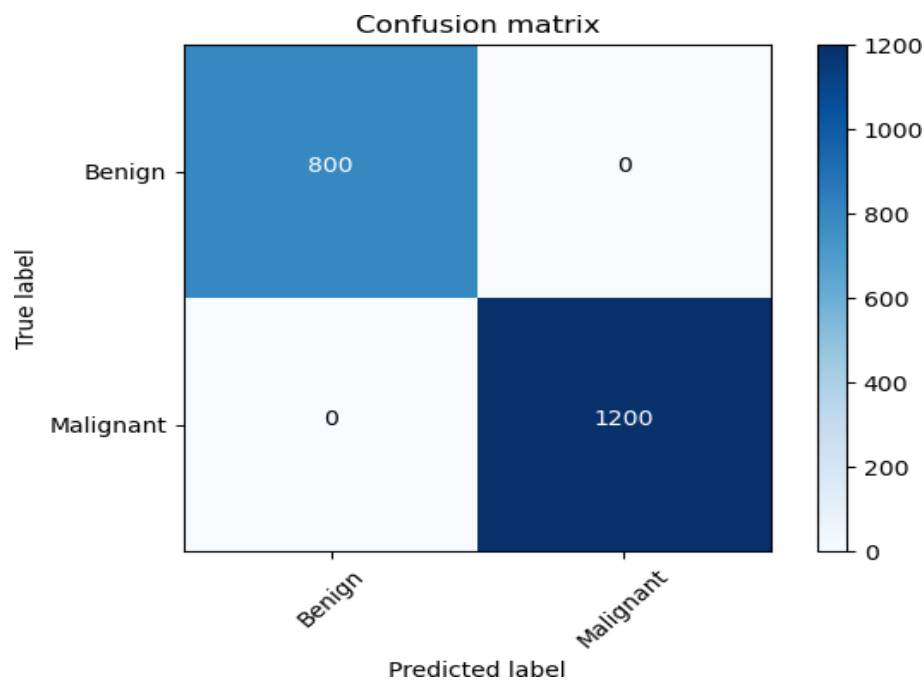
1. Base Model: The EfficientNetB0 model is loaded as the base model. This pre-trained model contains several convolutional layers that have learned to extract meaningful features from images.
2. Flatten Layer: A flatten layer is added to convert the output of the convolutional layers into a 1-dimensional feature vector.
3. Dense Layers: Two dense layers are added after the flatten layer. Each dense layer consists of 4096 units and uses the ReLU activation function. The dense layers help in learning complex patterns and representations from the extracted features.
4. Dropout Layers: Dropout layers are inserted after each dense layer to prevent overfitting. Dropout randomly drops a fraction of the input units during training, which helps in reducing the model's dependence on specific features.
5. Output Layer: The final dense layer contains num_classes units and uses the softmax activation function. It produces the probability distribution over the classes, indicating the likelihood of the input image belonging to each class.

With a balance between efficiency and performance, the EfficientNetB0 model is a potent option for picture categorization applications. Its architecture and pre-training on ImageNet provide it the ability to take advantage of learnt representations and extract significant features from input photos. The model may be customized on certain datasets to increase performance, making it a helpful tool for a variety of picture categorization applications.

The confusion matrix provides insights into the model's performance, showcasing the true positive and true negative rates. The visualization of correctly classified images further confirms the model's ability to accurately differentiate between benign and malignant breast cancer cases.

	precision	recall	f1-score	support
0	0.67	1.00	0.80	2
1	0.00	0.00	0.00	1
2	0.67	0.67	0.67	3
Accuracy			0.67	6
Macro avg	0.44	0.56	0.49	6
Weighted avg	0.56	0.67	0.60	6

Accuracy on whole data: 0.498753



The compound scaling method used by EfficientNet, which strikes an acceptable balance between model complexity and performance, turns out to be an effective method for medical picture classification tasks. EfficientNet shows its ability to increase diagnostic accuracy while preserving computational efficiency by utilizing pre-trained models and including customization. The architecture and optimization of the model make it a fantastic option for healthcare practitioners looking to integrate deep learning into their workflows for diagnosing breast cancer.

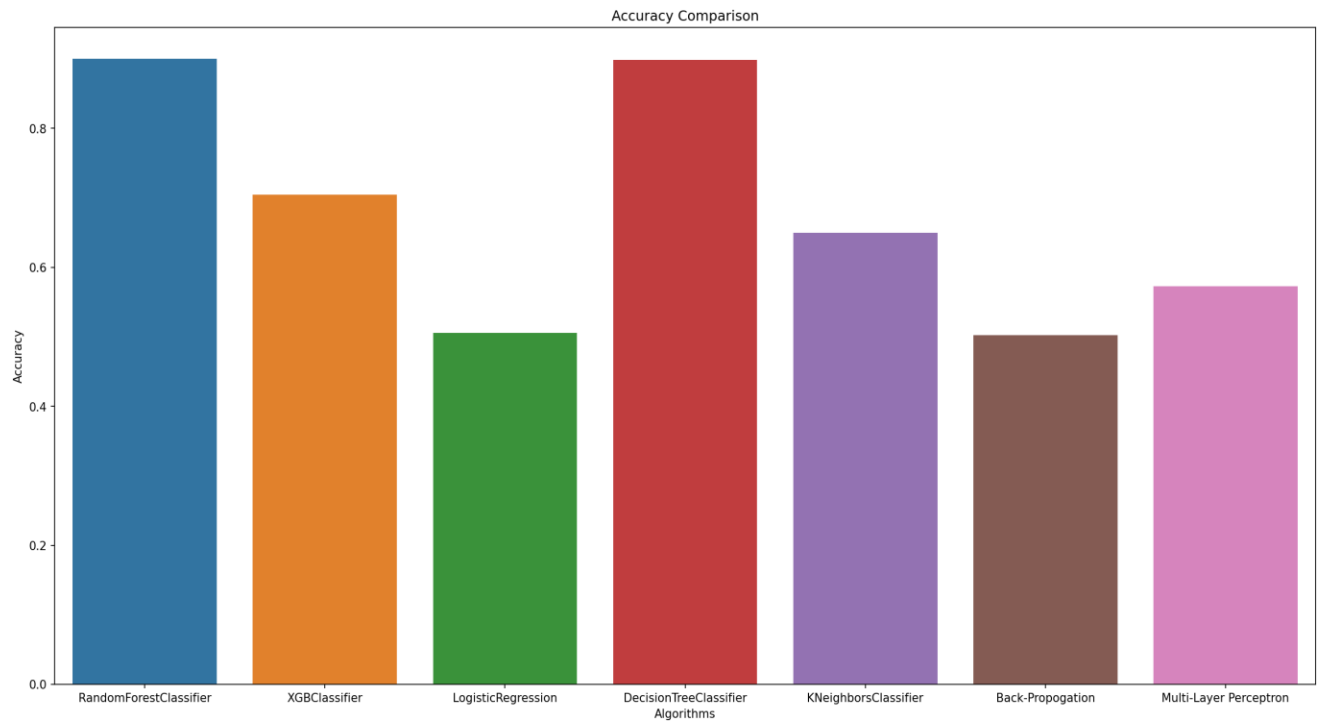
7. RESULT ANALYSIS

- M ---> Malignant (represents persons who is having Sevier stage of cancer)
- B ---> Benign (represents persons who are having starting stage)

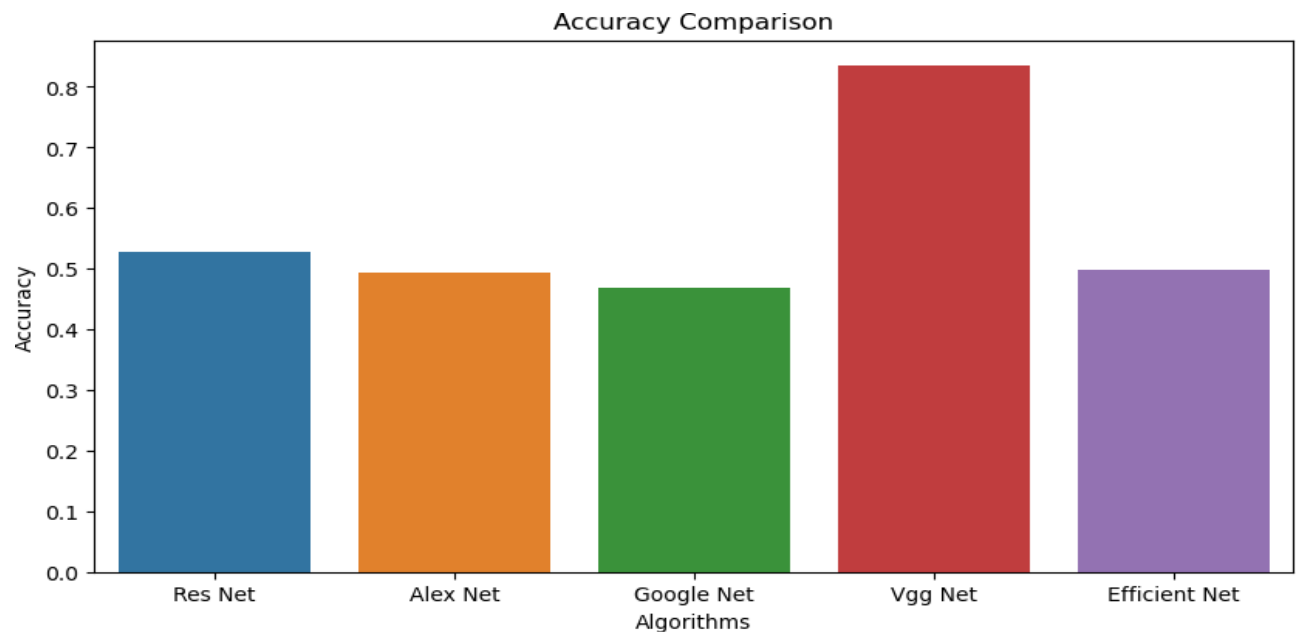
On the basis of accuracy, we examined and compared all of the algorithms. We used popular assessing criteria such as Precision, Recall, F-Score/F-Measure. Summarizes the outcomes of all algorithms in terms of accuracy, F-Score/F-Measure, precision, and recall. We can observe that the Logistic Regression and Back propagation have predicted nearly the same in terms of Accuracy which are 50, XGBoost has given 70%, Decision tree has given 89.8%, Multi-Layer Perceptron and K-Nearest Neighbor has given 57.2 and 64.9 as well as Random forest has given 90.001. In terms of accuracy and other evaluation metrics, Random Forest outperformed all other algorithms. So, according to the results obtained, We came to the conclusion that the Random Forest algorithm accurately predicted the best outcome. For Deep Learning models AlexNet and EfficientNet has shown the same Accuracy of 49, ResNet has shown 52, GoogleNet has given the least accuracy of the models which is 46 and VGG has given the highest accuracy of 83%. VGGNet surpassed all other algorithms in terms of accuracy and other evaluation parameters. So as per our project and our calculations we got the highest Accuracy Random Forest Algorithm that is 90 compared to all the machine learning algorithms that we have used. While coming to deep learning VGGNet gives the highest accuracy than other deep learning algorithms.

Comparison between Algorithms

The Accuracy comparison of all the implemented Machine Learning algorithms in graphical representation:



The Accuracy comparison of all the implemented Deep Learning algorithms in graphical representation:



8. CONCLUSION

Summary and Explanation

In summary, machine learning and deep learning algorithms have become effective tools for categorizing breast cancer. We can analyze enormous amounts of data and find detailed patterns by utilizing these cutting-edge methods that would not be visible to human experts alone. This capacity has the potential to completely change breast cancer treatment and early detection. These algorithms can develop the ability to distinguish between different tumor types and correctly forecast whether a case is benign or malignant by receiving training on extensive datasets of documented breast cancer cases. This has significant effects on increasing diagnostic precision and facilitating prompt interventions. It is vital to stress that machine learning and deep learning algorithms are aids, not replacements, for medical expertise, even while they offer healthcare practitioners valuable help. Qualified healthcare professionals should always have the last say in diagnostic and treatment choices. For the purpose of improving the precision, dependability, and generalize ability of breast cancer classification models, ongoing research and development in this area is essential. In the fight against breast cancer, we can pave the road for earlier detection, individualized treatment strategies, and better patient outcomes by utilizing the promise of machine learning and deep learning. Based on our research we found that Random Forest gives higher accuracy for the breast cancer classification using machine learning and VGGNet gives higher accuracy for breast cancer classification using deep learning.

9. FUTURE SCOPE

Future study for breast cancer classification using machine learning includes investigating the use of algorithms which have shown ability in improving the accuracy value of breast cancer classification. Another future work includes creation of customized breast cancer risk classification model based on machine learning algorithm which helps women in identifying the high risk of breast cancer and allows early detection. Multiple machine learning models can be combined to improve classification performance using ensemble learning approaches like random forests and gradient boosting. Ensemble approaches are able to increase accuracy, robustness, and generalization capabilities by utilizing the advantages of several models.

In order to present a complete picture of breast cancer, deep learning algorithms can combine different data modalities, including mammography, MRI, ultrasound, and genetics. Integrating many data sources can increase classification accuracy and provide for a more thorough picture of the disease. Successful treatment of breast cancer depends on early detection. Deep learning models can assist in the detection of tiny anomalies in medical imaging, assisting in the early detection of breast cancer. Deep learning algorithms can increase the sensitivity and specificity of early detection techniques by utilising vast datasets and sophisticated structures.

10. REFERENCES

- [1] Amrane, M., Oukid, S., Gagaoua, I. and Ensari, T., 2018, April. Breast cancer classification using machine learning. In 2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT) (pp. 1-4). IEEE.
- [2] Rane, N., Sunny, J., Kanade, R. and Devi, S., 2020. Breast cancer classification and prediction using machine learning. *International Journal of Engineering Research and Technology*, 9(2), pp.576-580.
- [3] Akbugday, B., 2019, October. Classification of breast cancer data using machine learning algorithms. In *2019 Medical technologies congress (TIPTEKNO)* (pp. 1-4). IEEE.
- [4] Michael, E., Ma, H., Li, H. and Qi, S., 2022. An optimized framework for breast cancer classification using machine learning. *BioMed Research International*, 2022.
- [5] Chugh, G., Kumar, S. and Singh, N., 2021. Survey on machine learning and deep learning applications in breast cancer diagnosis. *Cognitive Computation*, pp.1-20.
- [6] Jannesari, M., Habibzadeh, M., Aboulkheyr, H., Khosravi, P., Elemento, O., Totonchi, M. and Hajirasouliha, I., 2018, December. Breast cancer histopathological image classification: a deep learning approach. In *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)* (pp. 2405-2412). IEEE.
- [7] Liew, X.Y., Hameed, N. and Clos, J., 2021. An investigation of XGBoost-based algorithm for breast cancer classification. *Machine Learning with Applications*, 6, p.100154.
- [8] Rane, N., Sunny, J., Kanade, R. and Devi, S., 2020. Breast cancer classification and prediction using machine learning. *International Journal of Engineering Research and Technology*, 9(2), pp.576-580.

- [9] Tsang, J. and Tse, G.M., 2020. Molecular classification of breast cancer. *Advances in anatomic pathology*, 27(1), pp.27-35.
- [10] Sewak, M., Vaidya, P., Chan, C.C. and Duan, Z.H., 2007, August. SVM approach to breast cancer classification. In *Second international multi-symposiums on computer and computational sciences (IMSCCS 2007)* (pp. 32-37). IEEE.
- [11] Karthik, S., Srinivasa Perumal, R. and Chandra Mouli, P.V.S.S.R., 2018. Breast cancer classification using deep neural networks. *Knowledge Computing and Its Applications: Knowledge Manipulation and Processing Techniques: Volume 1*, pp.227-241.
- [12] Dai, X., Cheng, H., Bai, Z. and Li, J., 2017. Breast cancer cell line classification and its relevance with breast tumor subtyping. *Journal of Cancer*, 8(16), p.3131.
- [13] Zhang, Y.D., Satapathy, S.C., Guttery, D.S., Górriz, J.M. and Wang, S.H., 2021. Improved breast cancer classification through combining graph convolutional network and convolutional neural network. *Information Processing & Management*, 58(2), p.102439.
- [14] Venkatesan, E.V. and Velmurugan, T., 2015. Performance analysis of decision tree algorithms for breast cancer classification. *Indian Journal of Science and Technology*, 8(29), pp.1-8.
- [15] Hosni, M., Abnane, I., Idri, A., de Gea, J.M.C. and Alemán, J.L.F., 2019. Reviewing ensemble classification methods in breast cancer. *Computer methods and programs in biomedicine*, 177, pp.89-112.
- [16] Veronesi, U., Zurrida, S., Viale, G., Galimberti, V., Arnone, P. and Nolè, F., 2009. Rethinking TNM: a breast cancer classification to guide to treatment and facilitate research. *The breast journal*, 15(3), pp.291-295.
- [17] Wu, J. and Hicks, C., 2021. Breast cancer type classification using machine learning. *Journal of personalized medicine*, 11(2), p.61.
- [18] Rajaguru, H. and SR, S.C., 2019. Analysis of decision tree and k-nearest neighbor algorithm in the classification of breast cancer. *Asian Pacific journal of cancer prevention: APJCP*, 20(12), p.3777.
- [19] Paulin, F. and Santhakumaran, A., 2011. Classification of breast cancer by comparing back propagation training algorithms. *International Journal on Computer Science and Engineering*, 3(1), pp.327-332.
- [20] Bardou, D., Zhang, K. and Ahmad, S.M., 2018. Classification of breast cancer based on histology images using convolutional neural networks. *Ieee Access*, 6, pp.24680-24693.

- [21] Sahu, B., Mohanty, S. and Rout, S., 2019. A hybrid approach for breast cancer classification and diagnosis. *EAI Endorsed Transactions on Scalable Information Systems*, 6(20).
- [22] Zubair, M., Wang, S. and Ali, N., 2021. Advanced approaches to breast cancer classification and diagnosis. *Frontiers in Pharmacology*, 11, p.632079.
- [23] Viale, G., 2012. The current state of breast cancer classification. *Annals of oncology*, 23, pp.x207-x210.