BREAST CANCER CLASSIFICATION AND SEGMENTATION USING MACHINE LEARNING CLASSIFIERS AND U-NET

Susmitha Jejigari
Missouri S&T
Rolla, 65401
sjhtq@umsystem.edu

Niharika Gurram
Missouri S&T
Rolla, 65401
ngtg8@umsystem.edu

Rohith Reddy Yarramreddy
Missouri S&T
Rolla, 65401
ryfm5@umsystem.edu

Jayanth Vajja
Missouri S&T
Rolla, 65401
Jvnb5@umsystem.edu

Rishi Madhur Garimella
Missouri S&T
Rolla, 65401
rgc2f@umsystem.edu

Vikas Kumar Reddy Sareddy
Missouri S&T
Rolla, 65401
vsmw5@umsystem.edu

Abstract

Cancer is a condition in which cells in the body begin to grow uncontrollably. Among cancers, breast cancer is the most common in women in the United States. Over time, fewer individuals die from it, but it remains the second largest cause of cancer death in women. Fortunately, when detected early, it can be successfully treated in 70-80% of instances. In our research work we had done both classification and segmentation of breast cancer. For the classification purpose we applied machine learning classifiers such as SVM, LR, KNN, Naive Bayes, Light GBM, and ADA boost to labeled data comprising various attributes derived from breast tissue samples. In addition, we used the U-Net neural network design for segmentation, which allowed us to precisely identify malignant spots in mammogram images.

1. Introduction

Breast cancer is a devastating disease that affects hundreds of thousands of people every year. It occurs when cells in the breast tissue undergo malignant transformation, resulting in the creation of tumors. While breast cancer can strike at any age, the risk rises dramatically as women become older, with most of occurrences happening in those over 50. Early detection by routine screening procedures such as mammograms and breast self-exams is critical, as it increases the likelihood of successful treatment and survival.

Breast cancer is a huge global health issue that affects millions of people each year, regardless of gender. In the United States alone, It is estimated that one out of every eight women may get dense breast cancer at some time in her life. Surprisingly, men are not immune, with roughly 2,650 new cases reported each year. When detected early and before spreading beyond the breast, the 5-year relative survival rate is an amazing 99%. However, a delayed diagnosis might cause the cancer to spread and metastasis, making treatment more difficult.

Fortunately, we've made significant progress in the fight against breast cancer. We have made great progress in diagnosing, treating, and improving survival rates because of continuing research and technological improvements. Screening tools including mammography, ultrasound, and MRI scans have played a vital part in this improvement, helping doctors to detect breast cancer at earlier stages when it is simpler to treat.

However, effectively diagnosing and defining breast cancer cells inside breast tissue remains a significant challenge. Here is where machine learning (ML) comes into play. ML systems can scan large volumes of medical imaging data, detecting tiny patterns and traits that indicate malignant growths. Using these algorithms, healthcare providers can improve diagnostic accuracy, resulting in more customized and successful treatment options. Our project focuses especially on breast cancer classification and segmentation.

1.1. Breast Cancer Classification

Breast cancer classification starts with thorough data preparation, which is an important step in refining the raw information retrieved from breast tissue samples. This procedure consists of a series of changes and modifications designed to improve the dataset's quality and usability, allowing for more successful machine learning model training.

Initially, raw data retrieved from breast tissue samples may have errors, noise, or missing values. Data pretreatment aims to eliminate these flaws, resulting in a clean and uniform dataset suitable for robust model training. Data cleaning, missing value imputation, and outlier detection and removal are used to improve the dataset's integrity. Furthermore, feature extraction is vital in data preprocessing. Various features are retrieved from raw breast tissue samples to capture relevant information such as texture, shape, and intensity. These attributes serve as the foundation for training machine learning classifiers, allowing them to distinguish between malignant and non-cancerous regions in the data.

Following rigorous preprocessing and feature extraction, skilled doctors precisely label the dataset. Each sample is classified as "cancerous" or "non-cancerous," indicating whether it is benign or malignant.

We employed machine learning classifiers like SVM, LR, KNN, Naive Bayes, Light GBM and ADA boost on labeled data containing various features extracted from breast tissue samples.

SUPPORT VECTOR MACHINE

SVMs are a popular and basic machine learning technique. They belong to the supervised learning group and can handle both classification and regression tasks. SVM is especially good at solving complex learning problems like pattern recognition, bioinformatics, and medical diagnosis. It can work with separable datasets in both linear and nonlinear formats, making it extremely versatile.

There are two types of SVM: linear SVM (also known as simple SVM) and nonlinear SVM (or kernel SVM). Decision planes are important in SVM, which uses hyperplanes for classification. Decision boundaries are used to segregate data, and the region closest to the borders is referred to as the margin. SVM seeks to maximize the margin by achieving the greatest feasible separation between decision groups.

LOGISTIC REGRESSION

Logistic regression, a statistical workhorse, calculates the likelihood of events by fitting data to a mathematical function. It excels at categorizing objects, as opposed to linear regression, which works with continuous numbers. This supervised learning method excels in predicting binary outcomes (yes or no) by examining the relationship between several independent variables and a single dependent variable. In healthcare, logistic regression can use clinical and demographic data to determine a patient's immediate prognosis. While its primary output is a probability (which is frequently used for binary classifications with a 50% threshold), logistic regression also gives other information to aid decision-making.

K-NEAREST NEIGHBOR

KNN is a robust ML technique used to predict categories or values based on similarities between nearby data points. It works by identifying the nearest neighbors of a new data point and making predictions based on their majority class or average value. KNN is suited for a wide range of applications because it makes no assumptions about the underlying data pattern. However, its performance is dependent on parameters such as neighbor selection and distance metrics, and it can be computationally costly for large datasets. KNN is simple but effective, particularly when dealing with nonlinear data or complex patterns. However, it may encounter difficulties when dealing with high-dimensional data or areas with varying data density. Despite its simplicity, KNN is nevertheless an important tool in machine learning, providing a practical technique for creating predictions based on similarity with existing data points.

NAIVE BAYES

Naive Bayes is a machine learning algorithm that excels in categorizing objects. It operates by using Bayes' theorem, a statistical technique for determining the likelihood of an event occurring given current conditions. The term "naive" relates to the notion that qualities, or attributes, are independent of one another. For example, if an email is spam, the existence of certain words (such as "free" or "urgent") will not influence the appearance of others (such as misspellings or unusual formatting).

Despite this reduction, Naive Bayes frequently outperforms in real-world tasks, particularly when dealing with text classification. It is a popular choice for spam filters and sentiment analysis tools due to its speed, efficiency, and low training data requirements.

The algorithm calculates the probability of each category (such as "spam" or "not spam") depending on the features provided (words in an email). It then selects the category with the highest likelihood. Naive Bayes is especially useful for dealing with a large number of features because it analyzes each one independently. However, the assumption of independence can be limiting in cases where traits are genuinely connected. Overall, Naive Bayes is a powerful and intuitive algorithm that has proven useful in a variety of real-world situations. Its simplicity and efficiency make it an attractive option for a variety of categorization jobs.

LIGHT GBM

Light GBM, also known as Light Gradient Boosting Machine, is a sophisticated machine learning technique that excels at handling huge datasets with efficiency and speed. It belongs to the boosting algorithm family, which iteratively improves weak learners' predicting accuracy by focusing on previously misclassified examples. Light GBM is distinguished from classic gradient boosting methods by employing a novel methodology known as GOSS and EFB, which considerably expedite the training process while maintaining good accuracy. This makes Light GBM ideal for applications like classification, regression, and ranking, where speed and accuracy are critical factors.

ADAPTIVE BOOSTING

AdaBoost, or Adaptive Boosting, is a ML technique that combines numerous weak learners to produce a powerful classifier. Each weak learner is often a simple decision tree that performs marginally better than random guessing on a subset of the training data. AdaBoost iteratively trains these weak learners, assigning extra weight to data points that were incorrectly identified in prior iterations. This allows succeeding weak learners to focus more on difficult-to-classify cases, thereby boosting overall classification performance. Finally, AdaBoost combines all weak learners' predictions via a weighted majority vote, resulting in a robust and accurate classifier capable of handling complicated datasets.

1.2. Breast Cancer Segmentation

In the fight against breast cancer, early and accurate identification is essential. Mammography is essential in this process, but accurately identifying tumors within these pictures necessitates a critical step: segmentation.

This procedure entails painstakingly drawing the boundaries of malignant spots. U-Net, a deep learning architecture, stands up as an effective tool for this endeavor.

U-Net

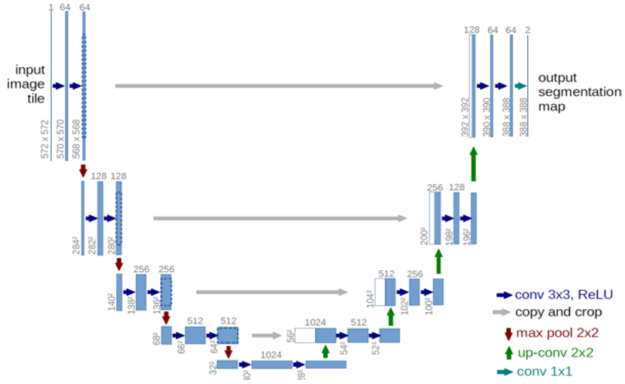


Figure 1: U-Net Architecture[6]

U-Net's secret is its two-pronged approach. The first component, the contracting path, methodically studies the mammography to gain a comprehensive understanding of the breast tissue. The second element, the extending path, refines this information, enabling for high-precision detection of malignant spots. By feeding a mammogram into U-Net, we get a segmentation mask. This mask labels each pixel to indicate whether it is in a malignant zone or healthy tissue. The benefits of segmentation are extensive. It enables doctors to accurately analyze the size, shape, and spread of malignancies. Armed with this information, they can devise the most effective treatment strategy, including selecting the right surgical margins. U-Net's strength is its ability to overcome the problems inherent in mammography images. These photos frequently show low contrast, noise, and changes in breast density. U-Net, on the other hand, is specifically intended to deal with such difficulties.

2. Related Works

In this paper [1] by Motamed et al. studied the use of machine learning to predict breast cancer. Their research of over 5,000 patients found that Random Forest had the most promising results. This algorithm has an accuracy of 80%, which means it accurately recognized breast cancer in 8 out of 10 cases. Furthermore, it achieved an AUC of 0.56, indicating that it reduced false positives. Gradient Boosting also performed well, with an AUC of 0.59. Importantly, the study discovered that integrating mammographic data increased model performance, highlighting its usefulness in machine learning-based breast cancer prediction.

In a study [2] on machine learning for breast cancer categorization, Meriem Amrane looked at two algorithms: Naive Bayes (NB) and K-Nearest Neighbors. Their investigation produced promising results, with KNN outperforming other models. Notably, KNN achieved a remarkable accuracy of 97.51%, indicating its efficacy in detecting breast cancer. This study demonstrates the potential of KNN as a useful tool for accurate breast cancer classification using machine learning.

This study [3] examined a cutting-edge technique to breast cancer categorization that produced outstanding results on the BUSI dataset. Their strategy achieved 90% accuracy by combining sophisticated deep learning models (Inception V3, ResNet50, and DenseNet121) within a meta-learning framework. This novel approach resulted in better generality and accuracy, particularly for malignant tumor diagnosis. The study emphasizes the potential of meta-learning and ensemble techniques to improve breast cancer diagnosis.

Another study [4] looked at a combined approach to automated breast cancer diagnosis utilizing ultrasound imaging. This strategy, like ours, focuses on early detection while taking advantage of ultrasound's capabilities. Their methodology used a multiscale classification model to categorize cancer and a U-shaped DDA-Net segmentation model with dual attention to precisely define tumor boundaries. Interestingly, they chose the whale optimization approach for feature selection in classification. Their results were outstanding, with high segmentation Dice coefficients (more than 87%) and classification accuracy surpassing 97% with strong precision. This enhances the promise of our strategy, which employs machine learning and picture segmentation for reliable and accurate early-stage breast cancer diagnosis via ultrasound imaging.

3. Proposed Method

Our project, "Breast Cancer Classification and Segmentation," provides a machine learning-based approach to automated breast cancer detection. This strategy tries to enhance early detection accuracy, which may lead to better patient outcomes.

Classification for Cancer Identification:

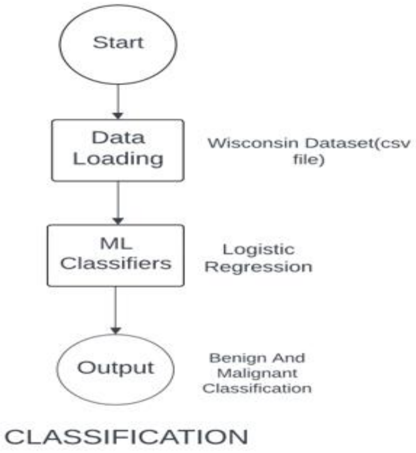


Figure 2: Block Diagram for Classification

Data Acquisition and Preprocessing: We use labeled breast cancer data to train and evaluate our classification models. This data goes through preprocessing steps to assure quality and consistency, which may include normalization, noise reduction, or image resizing.

Classifier Selection and Training: We investigate multiple machine learning classifiers to measure their usefulness in detecting breast cancer. Our initial analysis employs Logistic Regression (LR), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naive Bayes, AdaBoost, and Light GBM. Each of these algorithms has distinct benefits and may perform differently depending on the data qualities.

Performance Evaluation: We use metrics such as accuracy, Area Under the Curve (AUC), and Receiver Operating Characteristic (ROC) to evaluate the performance of each classifier. Based on these characteristics, we select the classifier that makes the most accurate and reliable difference between malignant and healthy tissue samples in our dataset.

Our evaluation procedure, which included criteria such as accuracy, Area Under the Curve (AUC), and Receiver Operating Characteristic (ROC), determined that Logistic Regression was the most successful classifier for our specific dataset. This means that, for the data we used and the evaluation measures we specified, Logistic Regression performed admirably at distinguishing between malignant and healthy tissue samples. This finding demonstrates the applicability of Logistic Regression for this specific classification problem in our project.

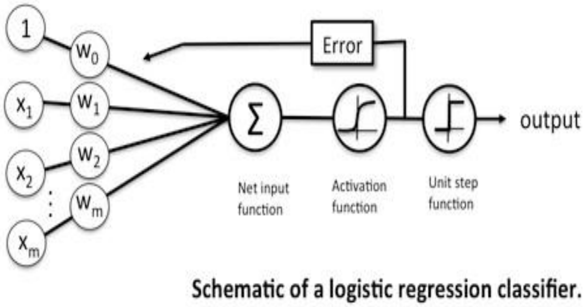
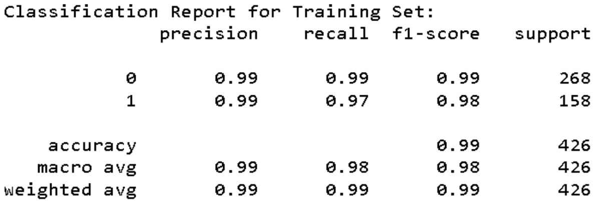


Figure 3:Architecture of LR



```
Classification Report for Training Set:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       268
           1       0.99      0.97      0.98       158

    accuracy                           0.99       426
   macro avg       0.99      0.98      0.98       426
weighted avg       0.99      0.99      0.99       426
```

Figure 4: Classification Report for training set of LR

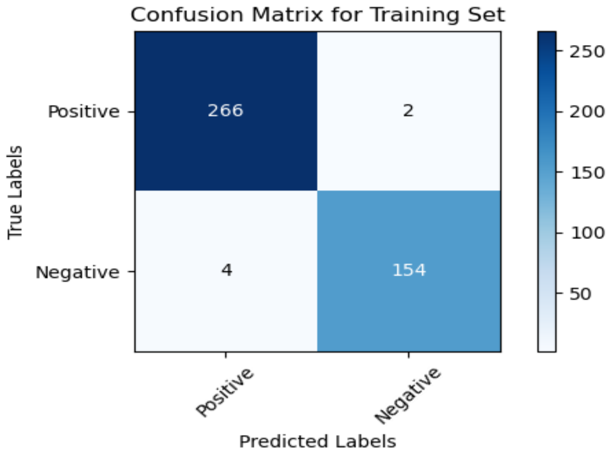


Figure 5: Confusion Matrix for Training Set of LR

Accuracy= (TP+TN)/(TP+TN+FP+FN)

Training Accuracy: 0.9859154929577465
Area under ROC curve: 0.9836104288683166

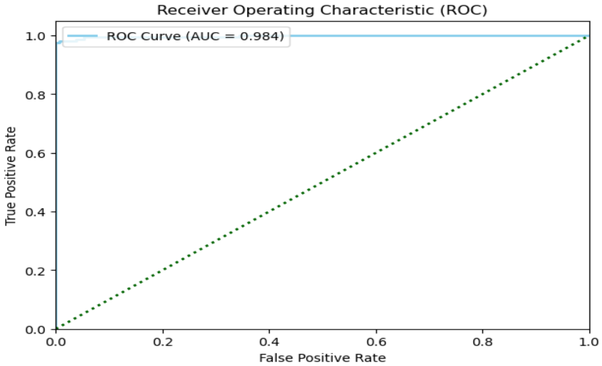
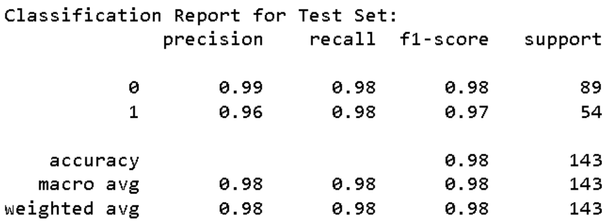


Figure 6: ROC Curve of training set of LR



```
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.99      0.98      0.98        89
           1       0.96      0.98      0.97        54

    accuracy                           0.98       143
   macro avg       0.98      0.98      0.98       143
weighted avg       0.98      0.98      0.98       143
```

Figure 7: Classification Report for test set of LR

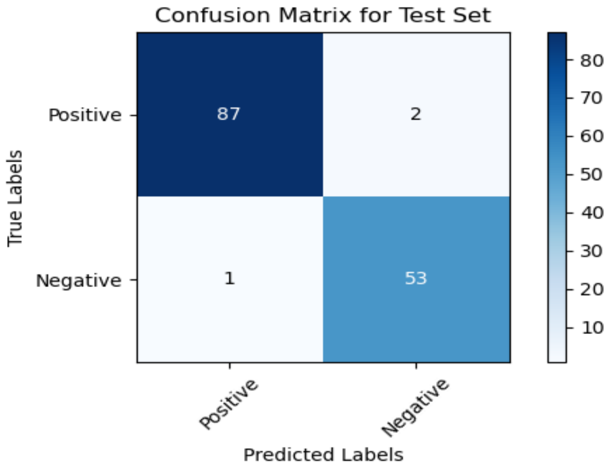


Figure 8: Confusion Matrix of Testing Set of LR

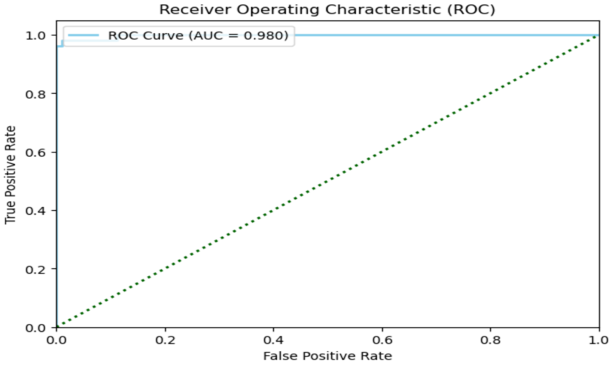


Figure 9: ROC Curve of test set of LR
Test Accuracy: 0.9790209790209791
Area under ROC curve: 0.9795047856845608

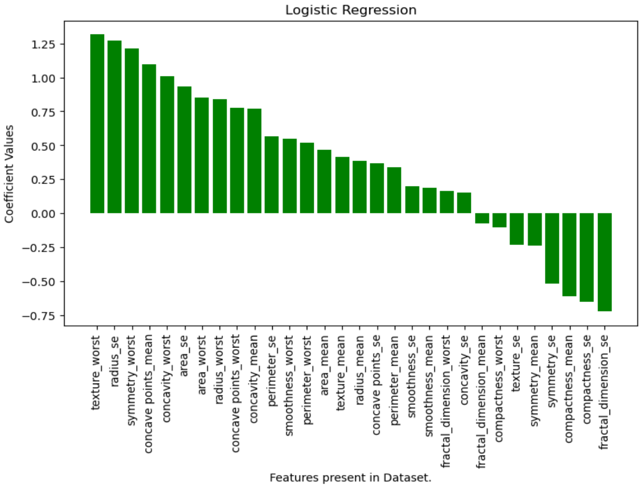


Figure 10: Features present in Data set

Segmentation for Precise Tumor Localization:

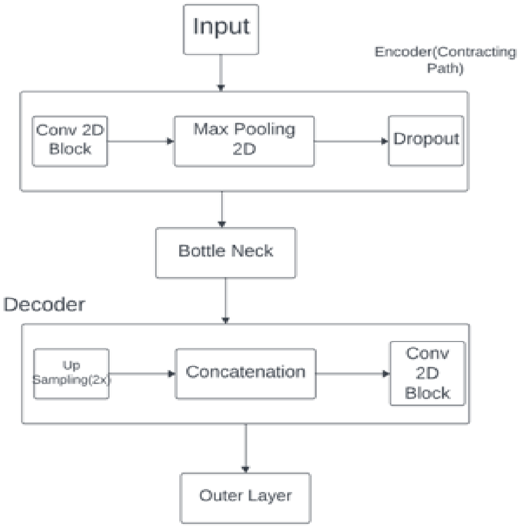


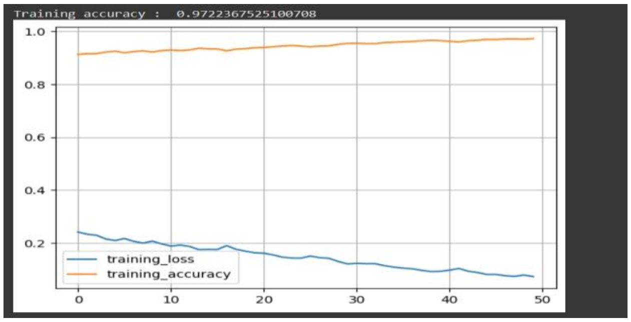
Figure 11: Block Diagram for Segmentation

Segmentation Model Selection: We use the U-Net architecture as the foundation of our segmentation model. U-Net's deep learning skills enable it to capture complicated relationships among medical pictures as well as delineate specific regions of interest, making it suitable for tumor mask predictions.

BUSI Dataset for Mask Generation: The BUSI dataset, which includes tagged breast cancer images, serves as the basis for training our segmentation model. By feeding these photos into the U-Net model, we hope to create exact tumor masks. These masks effectively highlight the specific regions of malignant tissue in the photos.

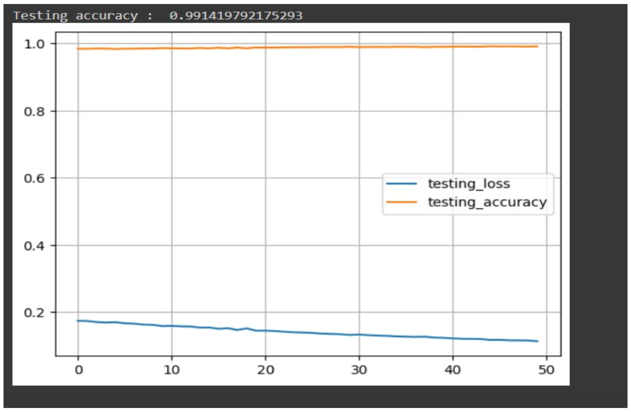
Model Training and Refinement: The U-Net model is trained using the BUSI dataset, gradually learning to identify and outline tumor areas with high accuracy. We may use approaches such as hyperparameter tuning to improve the model's performance and ensure that it generalizes well to new data.

Segmentation Model Training accuracy :0.97%
Segmentation Model Testing accuracy :0.99%



Training accuracy : 0.9722367525100708

Figure 12: Segmentation Training



Testing accuracy : 0.991419792175293

Figure 13: Segmentation Testing

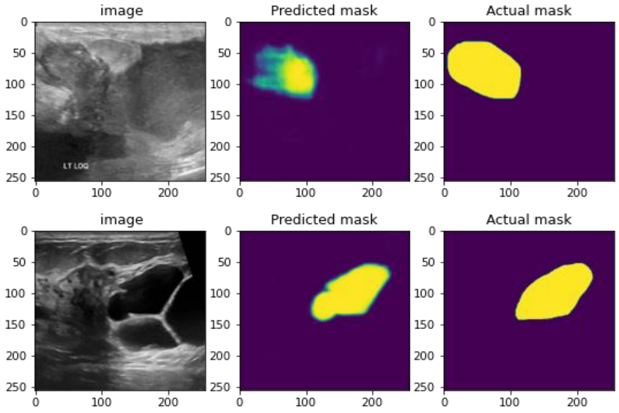
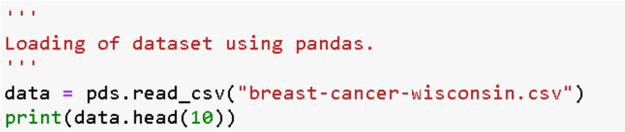


Figure 14:Segmentation of tumor

4. Experiments

For classification whether the tumor is “cancerous" or "non-cancerous," indicating whether it is benign or malignant breast cancer Wisconsin dataset is used.



```
'''
Loading of dataset using pandas.
'''
data = pds.read_csv("breast-cancer-wisconsin.csv")
print(data.head(10))
```

Figure 15:Dataset

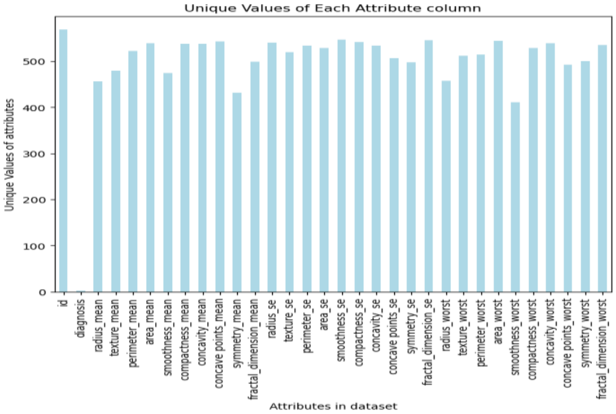


Figure 16: Attributes in dataset

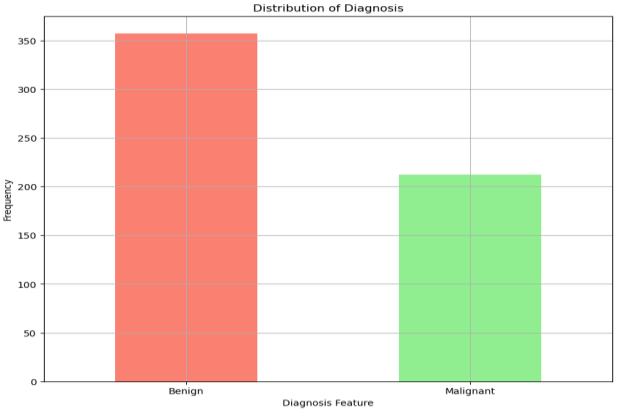
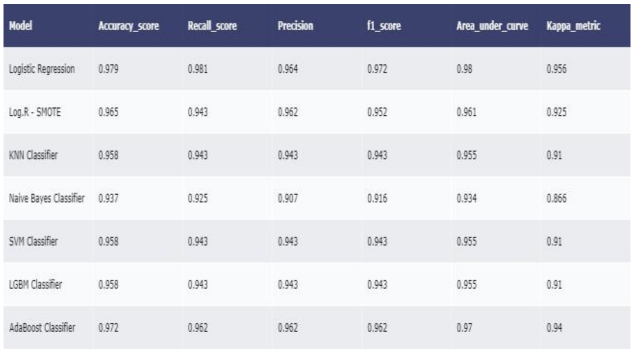


Figure 17: Number of malignant and benign count

We investigate multiple machine learning classifiers to measure their usefulness in detecting breast cancer. Our initial analysis employs LR, SVM, K-Nearest Neighbors , Naive Bayes, AdaBoost, and Light GBM.



Model	Accuracy_score	Recall_score	Precision	f1_score	Area_under_curve	Kappa_metric
Logistic Regression	0.979	0.981	0.964	0.972	0.98	0.956
Log.R - SMOTE	0.965	0.943	0.962	0.952	0.961	0.925
KNN Classifier	0.958	0.943	0.943	0.943	0.955	0.91
Naive Bayes Classifier	0.937	0.925	0.907	0.916	0.934	0.866
SVM Classifier	0.958	0.943	0.943	0.943	0.955	0.91
LGBM Classifier	0.958	0.943	0.943	0.943	0.955	0.91
AdaBoost Classifier	0.972	0.962	0.962	0.962	0.97	0.94

Figure 18: Comparison of metrics of different classifiers

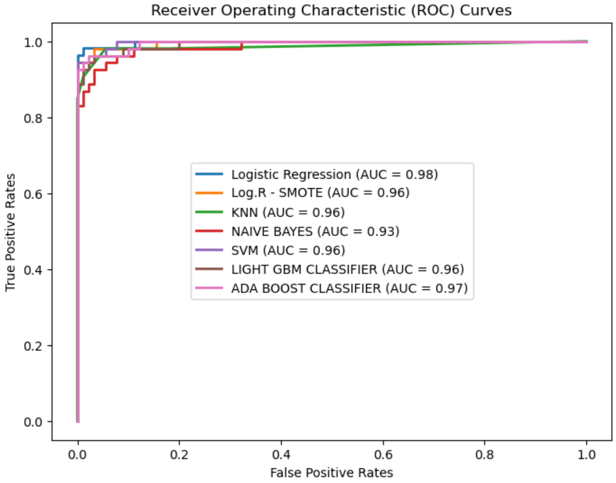


Figure 19: ROC curves of different classifiers

5. Conclusion

Our project, "Breast Cancer Classification and Segmentation," looked into machine learning's potential for automated breast cancer detection. We tested several classifiers (Logistic Regression, SVM, KNN, Naive Bayes, AdaBoost, and LightGBM) on labeled data. Notably, Logistic Regression demonstrated an excellent 98% accuracy with high AUC and ROC metrics, suggesting its superior capacity to distinguish between malignant and healthy tissues. We used the U-Net architecture on the BUSI dataset to forecast tumor masks and successfully identified malignant spots. Overall, this experiment demonstrates the promise of machine learning in classification and segmentation. The remarkable performance of Logistic Regression and the success of U-Net mask prediction pave the door for additional research into these techniques in order to produce more accurate and automated diagnostic tools.

References

[1] Reza Rabiei; Prediction of Breast Cancer using Machine Learning Approaches, 10.31661/jbpe.v0i0.2109-1403

[2] Amrane, Meriem; Oukid, Saliha; Gagaoua, Ikram; Ensari, Tolga (2018). [IEEE 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) - Istanbul, Turkey (2018.4.18-2018.4.19)] 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) - Breast cancer classification using machinelearning. (), 1–4. doi:10.1109/EBBT.2018.8391453

[3] Muhammad Danish Ali ; Adnan Saleem ; Breast Cancer Classification through Meta-Learning Ensemble Technique Using Convolution Neural Networks ,10.3390/diagnostics13132242

[4] Muhammad Sharif, Shui-Hua Wang; Breast cancer classification and segmentation framework using multiscale CNN and U-shaped dual decoded attention network, https://doi.org/10.1111/exsy.13192

[5] G. R. Jothilakshmi and A. Raaza, ‘‘Effective detection of mass abnormal- ities and its classification using multi-SVM classifier with digital mam- mogram images,’’ in Proc. Int. Conf. Comput., Commun. Signal Process. (ICCCSP), Jan. 2017, pp. 1–6.

[6] https://towardsdatascience.com/unet-line-by-line-explanation-9b191c76baf5

[7] K. Whitaker, ‘‘Earlier diagnosis: The importance of cancer symptoms,’’ Lancet Oncol., vol. 21, no. 1, pp. 6–8, Jan. 2020.