Breast Cancer Detection and Classification Using Ultrasound Images & Classic Machine Learning Algorithms

Abdallah Magdy
Systems and Biomedical Engineering
Cairo University

Muhammad Alaa
Systems and Biomedical Engineering
Cairo University

Nouran Mahmoud
Systems and Biomedical Engineering
Cairo University

Omar Emad
Systems and Biomedical Engineering
Cairo University

***Abstract*— This paper investigates breast cancer prediction and classification utilizing classic machine learning algorithms applied to ultrasound images. Three distinct approaches were explored, achieving notable results. Gradient Boosting and AdaBoost classifiers yielded F1 scores of 0.837 and 0.902, respectively, in distinguishing normal from cancerous tissue. Additionally, they demonstrated F1 scores of 0.87 and 0.85 in classifying benign versus malignant tumors. Radiomics-Based classification yielded F1 score of 0.92 for the same task. These findings suggest promising avenues for enhancing early detection and diagnostic accuracy in breast cancer assessment.**

Keyword—Machine learning, Medicine, Breast Cancer, Classification, Early Detection

I. Introduction

Breast cancer detection and classification is a critical area of medical research aimed at improving diagnostic accuracy and patient outcomes. This report presents three distinct approaches for detecting and classifying breast cancer using ultrasound images, evaluating their methodologies, experimental results, and overall effectiveness.

II. Theoretical Background

A. Cancer

Cancerous tumor results from the abnormal growth of cells that invade the surrounding tissues of the human body. There are two types of tumors, benign and malignant, and when no tumor is present in the breast, it is considered normal. Benign tumor cells are benign cells that only grow locally and cannot spread by invasion. While malignant tumors are cancer cells and could multiply out of control, spread to different parts of the body, and invade surrounding tissues.

B. Breast Cancer

Breast cancer is a disease in which cells in the breast grow out of control. There are different types of breast cancer. The type of breast cancer depends on which cells in the b east develop into cancer. Breast cancer can start in different parts of the breast. A breast consists of three main parts: lobules, milk ducts, and connective tissue. The lobules are the glands that produce milk. Ducts are tubes that carry milk to the nipple. Connective tissue (consisting of fibrous and fatty tissue) surrounds and holds everything together. Most breast cancers start in the ducts or lobules. Breast cancer can spread through blood vessels and lymphatic vessels outside the breast. When breast cancer spreads to other parts of the body, it is said to have metastasized.

C. Breast Cancer Statistics

The most commonly occurring cancer in women is breast cancer (BrC). As claimed by the World Health Organization (WHO), BrC was diagnosed in 2.3 million women, and 685,000 deaths were recorded globally in 2020. In addition, the WHO predicts that the number of new BrC patients will increase by seventy percent (70%) in the next twenty years. Besides, BrC is the 5th-most deadly disease out of distinct cancer types, such as lung, colorectal, liver, and stomach cancers. which further states that in 2030 the death ratio from cancer is expected to increase up to 27 million. So, on-time and accurate detection, early diagnosis, and active prevention are critical requirements for reducing mortality rate among women.

D. Breast Cancer in Egypt

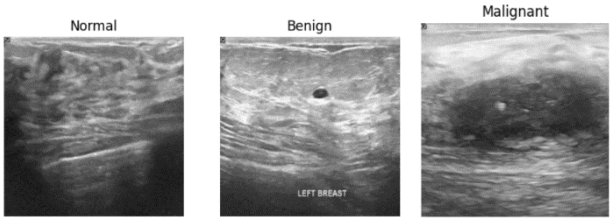
In Egypt, breast cancer is the most common malignancy in women, accounting for 38.8% of cancers in this population, with the estimated number of breast cancer cases nearly 22,700 in 2020 and forecasted to be approximately 46,000 in 2050. It is estimated that the breast cancer mortality rate is around 11%, being the second cause of cancer-related mortality after liver cancer.

E. Input Output Description

The input is ultrasound scans for breast cancer and the output is according to classification stage. If it's the first stage, the output is one of the two classes (normal, cancerous). If it's the second stage, the output is one of the two classes (benign, malignant).

III. Dataset Description

We are using the "Breast Ultrasound Images Dataset" (Dataset BUSI) [1] which is collected by Dr. Aly Fahmy. The data collected in baseline include breast ultrasound images among women in ages between 25 and 75 years old. It was collected in 2018. The number of patients is 600 female patients, and it consists of 780 images along with segmentation masks for the benign and malignant cancers. The data is categorized into three classes: Normal, benign, and malignant.



IV. Related Work

The field of breast cancer classification has been thoroughly examined in numerous scientific studies, utilizing a wide range of imaging modalities. These include mammography (which employs X-ray imaging), magnetic resonance imaging (MRI), sonography (which uses ultrasound imaging), and thermography. Our research, however, is specifically focused on the application of traditional machine learning techniques to ultrasound images. Existing research conducted can be broadly classified into three distinct categories:

1- Approaches that involve comprehensive preprocessing and feature extraction using handcrafted methods, followed by the application of a single classifier.

2- Techniques that employ deep learning (DL) for feature extraction, coupled with the use of ensemble classifiers.

3- Methods that involve minimal feature extraction, also combined with the use of ensemble classifiers.

In the first category, Mishra et al [1] used the same dataset. Their method was: 30-70 test split, creating RGB masked images, extracting hand crafted features, performing recursive feature elimination, performing synthetic minority oversampling technique (SMOTE) then training the models.

They reported an accuracy of 96.5% and f1-score of 94.8% using Gradient Boosting Classifier. Their approach is comprehensive and suitable for ultrasound data, but one killing weakness is that it's not reproducible as we achieved an accuracy, and f1-score are 85%. We have also sadad et al. [2] details a method for classifying breast cancer using our dataset.

It employs marker-controlled watershed transformation for accurate lesion segmentation – Which is weird given that they have the masks already – and extracts both shape-based and texture features. Three classifiers—Decision Tree, KNN, and RUSBoost—are used, with the ensemble method providing the best accuracy (96%). Strengths include effective feature capture, robust segmentation, and high accuracy, while weaknesses involve computational complexity, dependency on image quality, and the need for manual parameter tuning. The method's generalizability to other datasets also requires further validation.

In the second category Y. Eroğlu et al. [3] developed a hybrid CNN system using AlexNet, MobileNetV2, and ResNet50 to diagnose breast cancer lesions, classifying them as benign, malignant, or normal. By concatenating and selecting the most important features with the mRMR method, and then they used machine learning classifiers like SVM and KNN, they achieved the highest accuracy of 95.6% with the SVM classifier. This can be a good approach, especially that hand-crafted features need extensive domain knowledge, but it doesn't utilize the full power of CNNs in image classification.

In the thir A. Halder et al. [4] extracted ultrasound based radiomic features from the same dataset. They have been extracted from the tumor as well as the peritumor region and utilized several classifiers like SVM, Random Forest, Logistic Regression and XGBoost. SVM Classifier achieved 96% accuracy with them.

As for the stat-of-the-art model for breast cancer classification, it utilizes a robust encoder-decoder architecture for accurate lesion segmentation in ultrasound images. It employs a dual-staged feature fusion technique combining features from ResNet50V2, NASNetLarge, and EfficientNetB7, enhancing classification accuracy. Data augmentation using traditional techniques and CycleGAN addresses data scarcity. The model achieves a high accuracy of 99.7% and an AUC of 0.999, showcasing superior performance across multiple datasets. Despite its high computational complexity, this model's advanced techniques and robust evaluation metrics establish it as a state-of-the-art approach for breast cancer classification.

V. Data Preprocessing

In this section we discuss the problems we see in the data and the preprocessing steps to make the data ready for modeling

A. *Labeling and Data Loading*

The dataset was organized into directories corresponding to each class ('normal/', 'benign/', 'malignant/'). Images were loaded from these directories and labeled accordingly.

B. *Undersampling*

Since the dataset was unbalanced, we performed undersampling to balance it. We used only 133 images for each class (normal, benign, and malignant) to ensure uniform class distribution.

C. *Data Splitting*

The dataset was split into training and testing sets, with 90% of the data allocated for training and 10% for testing. This split was stratified to ensure proportional representation of each class in both sets.

D. *Normalization*

Pixel values of the images were normalized to the range [0, 1]. This preprocessing step ensures that the input data is on a consistent scale, aiding in faster convergence during model training.

E. *Resolution*

All ultrasound images were resized to 64x64 pixels (SIZE = 64). This resolution standardization ensures uniformity in input dimensions for the machine learning models.

F. The segmentation masks were provided for benign and malignant cancers.

VI. Methods

A. *Gradient Boosting Classifier*

Gradient Boosting is an ensemble learning technique that builds a strong predictive model by sequentially combining multiple weak learners, typically decision trees. Each subsequent tree corrects the errors made by the previous ones, with a focus on the regions where the previous models performed poorly.

The loss function for Gradient Boosting typically involves minimizing the negative gradient of a differentiable loss function, such as the logistic loss for classification problems or the squared loss for regression problems.

For binary classification, the logistic loss function is commonly used:

$$L(y, f(x)) = \sum_{i=1}^{N} log\left(1 + e^{-y_i f(x_i)}\right)$$

Where y_i is the true label (-1 or 1), $f(x_i)$ is the predicted output, and N is the number of samples.

B. *AdaBoost Classifier*

AdaBoost, short for Adaptive Boosting, is a boosting algorithm that sequentially fits a series of weak learners on different weighted versions of the data. It adjusts the weights of incorrectly classified instances to focus subsequent learners on harder examples. The final model combines the predictions of all weak learners (Logistic

regression in our case) through a weighted sum, giving more weight to the ones with higher accuracy.
The loss function for AdaBoost is exponential loss, which penalizes misclassifications exponentially.

$$L(y, f(x)) = \sum_{i=1}^{N} exp(-y_i f(x_i))$$

Where y_i is the true label (-1 or 1), $f(x_i)$ is the predicted output, and N is the number of samples.

VII. Experiments And Results

For our experiments, we employed three distinct machine learning approaches to classify breast ultrasound images into three categories: Normal, Benign, and Malignant. The dataset consisted of ultrasound images with resolutions standardized to 64×64 pixels. Here, we outline the methodologies, parameter choices, evaluation metrics, and results obtained from each approach.

In this study, we utilize several key metrics to evaluate the performance of our machine learning models in classifying breast tissue abnormalities. These metrics provide quantitative insights into the models' effectiveness in distinguishing between different classes and their overall predictive power.

A. Metrics

1) F1 Score

The F1 score is a metric that combines both precision and recall into a single value, providing a balanced assessment of a classifier's performance. It is particularly useful in scenarios where the class distribution is imbalanced.
The F1 score is calculated as the harmonic mean of precision (P) and recall (R): $F1 = 2x\frac{PxR}{P+R}$
Where:

- Precision $P = \frac{TP}{TP+FP}$ measures the accuracy of positive predictions.
- Recall $R = \frac{TP}{TP+FN}$ measures the proportion of actual positives correctly identified.

2) Confusion Matrix

The confusion matrix is a table used to describe the performance of a classification model. It shows the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each class, providing a detailed breakdown of correct and incorrect predictions.

3) False Negative Rate (FNR)

The false negative rate is the proportion of actual positives that are incorrectly identified as negatives by the model. It is calculated as follows:

$$FNR = \frac{FN}{FN + TP}$$

These metrics are essential in evaluating the diagnostic performance of machine learning models for breast cancer detection. By assessing the F1 score, analyzing the confusion matrix, and measuring the false negative rate, we can quantitatively measure the models' effectiveness in distinguishing between normal and abnormal breast tissue, aiding in clinical decision-making and patient care.

To solve the problem, we are talking about in this study, we have 3 different approaches, each approach with its drawbacks, we will mention them in the following part.

The hyperparameters for each model are chosen via 10-fold cross validation with grid search or Optuna auto tuner on the training set, and the test set was used to test the final performance and generalization to unseen data.

B. Hierarchical Models

Tackling the problem as a multi class classification problem was extremely difficult when limited to classical machine learning algorithms a lot of trails have been made using RF, SVMs along with principal component analysis but results were below 50% for the F1 score.

A better approach is to divide our multitask classification problem hierarchically into 2 binary classification problems. The first stage is to classify normal vs abnormal and the second stage for benign vs malignant.

C. Gradient Boosting Approach

A gradient boosting model was used with the training data and evaluated via 10-Fold cross validation, for the first model (Normal vs. Cancerous), the Mean F1 score was 0.87. For the second model (Benign vs. Malignant), the F1 score was 0.902. This approach provided a more accurate evaluation of model performance while maintaining data integrity, other trails with this approach involved dimensionality reduction using linear principal component analysis preserving 85, 90, 95 % of the explained variance but this did not seem to improve the results, in addition preprocessing like histogram equalization, contrast stretching and CLAHE were applied but did not seem to improve the results.

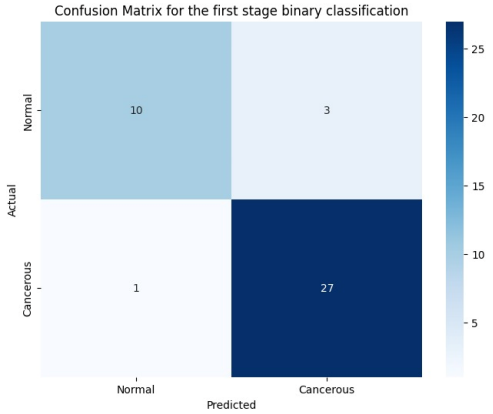


Figure 1: - Confusion Matrix for Normal vs Cancerous Model using Gradient Boosting

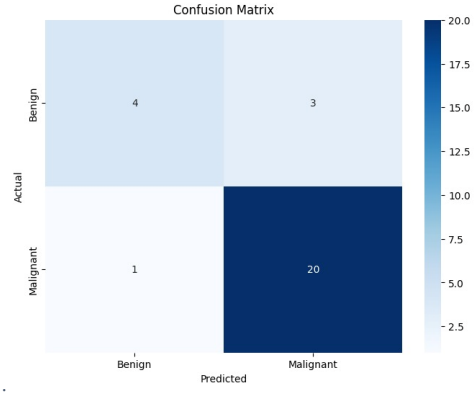


Figure 2 - Confusion Matrix for Bengin vs Malignant Model using Gradient Boosting.

D. *Adaptive Boosting*

In the third approach, we used AdaBoost with weak learners (Logistic Regression in our case) was used with the training data and evaluated via 10-Fold cross validation. Each classification task was treated independently. This approach allowed for optimizing models specifically for each classification challenge, enhancing overall accuracy.

The results demonstrated superior performance compared to the previous approaches. The F1 scores were 0.84 for Normal vs. Cancerous and 0.85 for Benign vs. Malignant, indicating excellent discriminative ability in both tasks.

The regularization hyper-parameter C for the logistic regression was selected via Optima auto tuner through 100 iterations and is found to be 0.001 with the best score.

By focusing on task-specific modeling, the hierarchical approach showcased the benefits of tailored solutions in complex medical diagnostics. Despite requiring more computational resources and methodological complexity, this approach significantly enhanced accuracy and provided insights into the effectiveness of customized models for specific clinical applications.

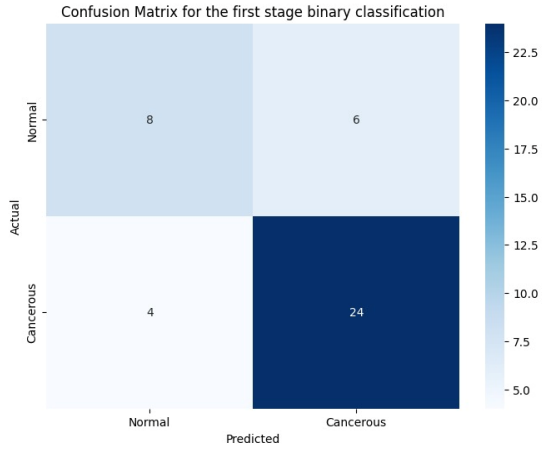


Figure 3- Confusion Matrix for Normal vs Cancerous Model using AdaBoost.

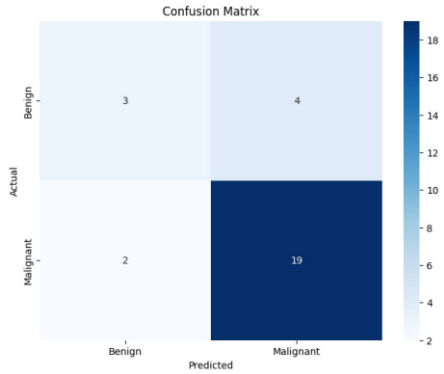


Figure 4 - Confusion Matrix for Benign vs Malignant Model using AdaBoost

Classification task was treated independently. This approach allowed for optimizing models specifically for each classification challenge, enhancing overall accuracy.

E. *Radiomic Based Approach*

In the original study [2], the authors developed a tumor classification model using ultrasound images. The initial data preprocessing involved combining images with corresponding masks to highlight tumor regions. Subsequently, shape, texture, and edge features were extracted to comprehensively represent tumors. Recursive Feature Elimination (RFE) was used for feature selection, and Synthetic Minority Over-sampling Technique (SMOTE) addressed class imbalance. The model, based on a Gradient Boosting Classifier, demonstrated robustness to variations in test size and feature selection while maintaining high performance. Notably, shape features (Hu moments), texture features (from gray-level co-occurrence matrix), and edge features (Shannon entropy from Canny edges) were useful for differentiating between benign and malignant tumors.

Our implementation to the paper approach yielded an F1 score of 0.85, while our modification to the paper approach performed a stable F1 score of 0.922 for classification between benign and malignant tumors.

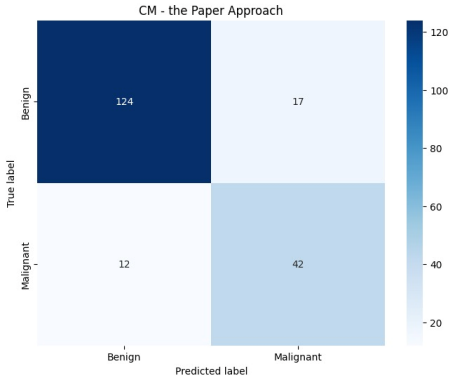


Figure 5- Confusion Matrix of Benign vs Malignant Model using the paper approach.

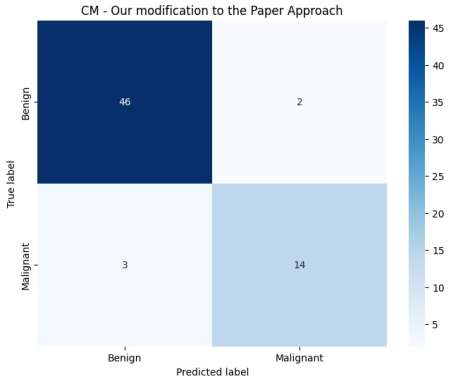


Figure 6 - Confusion Matrix of Benign vs Malignant using our modified paper approach.

VIII. CONCLUSION

Three distinct approaches were implemented and evaluated: a Gradient Boosting Classifier, an AdaBoost Classifier, and a Radiomic-Based Approach. As for the first stage, the best performing models were Logistic regression with adaptive boosting. It was able to outperform since the gradient boosting was overfitting and regularization did not improve its generalization performance. On the other hand, the logistic adaptive model seemed to respond to the regularization effect thus yielding better generalization performance. In the second stage, Radiomics based approach since it relied on utilizing segmentation masks and handcrafted features then recursive feature selection, thus having higher discriminating ability than the other 2 approaches.

IX. REFERENCES

[1] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," Data in Brief, vol. 28, p. 104863, Feb. 2020, doi: https://doi.org/10.1016/j.dib.2019.104863 .

[2] A. K. Mishra, P. Roy, S. Bandyopadhyay, and S. K. Das, "Breast ultrasound tumour classification: A Machine Learning—Radiomics based approach," Expert Systems, vol. 38, no. 7, May 2021, doi: https://doi.org/10.1111/exsy.12713 .

[3] T. Sadad et al., "Identification of Breast Malignancy by Marker-Controlled Watershed Transformation and Hybrid Feature Set for Healthcare," Applied Sciences, vol. 10, no. 6, p. 1900, Mar. 2020, doi: https://doi.org/10.3390/app10061900 .

[4] Y. Eroğlu, M. Yildirim, and A. Çinar, "Convolutional Neural Networks based classification of breast ultrasonography images by hybrid method with respect to benign, malignant, and normal using mRMR," Computers in Biology and Medicine, vol. 133, p. 104407, Jun. 2021, doi: https://doi.org/10.1016/j.compbiomed.2021.104407 .

[5] A. Halder, C. Bhowmick, P. K. Dutta and M. Mahadevappa, "Peri- and Intra-Tumoral Radiomic Signatures to Distinguish Benign and Malignant Tumors in Breast Ultrasound Images," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-5, doi: https://doi.org/10.1109/ICCCNT56998.2023.10306977 .

X. CHANGED PARTS IN PHASE 2

We've changed the problem totally compared to phase 1 as the data was mostly meaningless. In one of the experiments, we tested for the well-known distributions using Anderson Darling Test followed by imputing missing values then training 5 models (two of them are Gradient Boosting and Random Forest) with grid search and it didn't pass 75%-65% accuracy using proper methods.

As for PHASE I of this problem, using both linear and kernel PCA of different explained variance ratios and number of components, and simple hand-crafted features didn't improve the results, that's why we explored radiomics based features. Bagging Models also did not improve the results moreover it showed tendency to bias towards the most abundant class. Anomaly detection approach was observed using One Class SVM and it did not improve the results as well, that can be explained by the fact that the features are not good representations at the first place. SVMs showed higher variance even when regularized.

XI. FUTURE WORK

Deep learning approaches are the state of art for the majority of computer vision tasks so utilizing CNNs along with few-shot learning seems to be an outstanding approach. In addition, a multi-task meta learner can be trained and be used as a starting point not only for the classification task but also on segmenting the tumor once it was detected and localized.

XII. CONTRIBUTION

A. *Abdallah Magdy:*

1- Tried a multi class approach using RF, SVMs for images with features reduced using linear pca.
2- Tried the same models with Sift, Mops and HoGs features along with linear pca dimensionality reduction.
3- Tried some preprocessing like contrast stretching and histogram equalization.
4- Assisted with the hierarchical model approach with data preprocessing

B. *Omar Emad:*

1- Worked on a hierarchical model approach using gradient boosting.
2- Tried regularizing and tuning the above model.
3- Worked on a different Hand-Crafted Feature Extraction approach to enhance segmentation masks like snakes active contouring and Hough transforms to detect objects in the image.
4- Tried dimensionality reduction with linear PCA.

C. *Mohamed Alaa:*

1-Worked on Hierarchical models using adaptive boosting for logistic regression with its hyperparameter tuning.
2 - Tried lots of dimensionality reduction and manifold learning techniques but could not improve results.
3- Tried fusing some hand-crafted features but did not improve results.

D. *Nouran Mahmoud:*

1- Worked on the radiomics approach to classify benign vs malignant cancers.
2- Extensive literature review that led to extracting a lot of useful hand crafted features.
3- Trained and tuned a gradient boosting model along with smote resampling and recursive feature selection which improved the results.

XIII. FINAL BLOCK DIAGRAM