DDSM - Breast Classification via Machine Learning

Hayat Maya, Israeli Elor and Wasker Roy
Department of Computer Science
Ariel University

Abstract

This study investigates the feasibility of using machine learning models to classify breast tumors as benign or malignant based on radiologists' tabular data from the Digital Database for Screening Mammography (DDSM) dataset. The project prioritizes interpretable and efficient models while aiming to outperform the accuracy of traditional radiologist assessments.

We explored various classification algorithms and achieved promising results. The AdaBoost model achieved the highest accuracy (79.99%) and AUC score (0.8598) for the mass dataset using features like mass shape and margins, breast density, and tumor location. Including the radiologists' assessment scores further improved the accuracy to 83.1% and the AUC score to 0.9041, highlighting the value of combining human expertise with machine learning.

Extracting features from the Region of Interest (ROI) images within the DDSM dataset yielded mixed results. While features from calcification ROI images improved accuracy compared to tabular data alone, mass ROI features showed minimal impact.

The project highlights the potential of machine learning to assist in breast cancer diagnosis, but also reveals limitations due to missing data like patient demographics. Future work could involve incorporating additional patient information and exploring the integration of neural networks with the tabular data for potentially even more accurate and holistic results.

1 Introduction

Breast cancer is one of the most prevalent and recognized health challenges worldwide, affecting approximately 2.3 million women globally in 2022 and resulting in 670,000 reported deaths. Traditional pathology exams, while critical, can take between 7 and 10 days, causing significant stress for patients. The integration of machine learning into medical diagnostics can reduce human error, speed up diagnoses, and enhance classification accuracy. Our project aims to classify breast tumors as benign or malignant using mainly the tabular DDSM dataset written by the radiologist, focusing on interpretable and efficient models. By outperforming the radiologists' accuracy of traditional assessments, our work can help to better predict the tumor's pathology until the actual results.

The Digital Database for Screening Mammography (DDSM) dataset includes 3,567 entries, across 1,566 patients, containing both mass and calcification tumors. The dataset consists of tabular data and images and includes rather a lot of information about the tumor which can be explored. Each entry in the tabular data reflects a radiologist's observations about the tumor. Both the mass and calcification datasets contain patient IDs since some patients have multiple tumors, resulting in multiple entries. Additionally, most tumors are represented with two types of mammograms: Mediolateral Oblique (MLO) and Cranial-Caudal (CC) views. These views are important because the tumor's shape may appear different depending on the angle of the image. The dataset also includes breast tissue's density scores ranging from 1 to 4, specifies whether the tumor is in the left or right breast and provides the radiologist's assessment which follows the BI-RADS scale.

The BI-RADS assessment scores range from 0 to 6, but the DDSM dataset only includes scores from 0 to 5, as there are no cases of previously confirmed malignant tumors. Each score reflects a different level of suspicion, with 0 indicating an incomplete result that requires additional imaging. A score of 1 suggests a normal finding with no detected abnormalities, while 2 indicates a benign result. A score of 3 suggests a low probability of malignancy, with an estimated 5% chance. A score of 4 points to suspicious abnormalities, warranting a biopsy, and 5 indicates a high likelihood of malignancy usually above 95%. In addition to the radiologist's assessment, the dataset provides the final pathology result, which confirms whether the tumor is benign or malignant.

For the mass lesions, the dataset includes columns describing the mass shape and mass margins. The calcification dataset, on the other hand, includes the calcification distribution and type.

This study focuses on training machine learning models to accurately classify both mass and calcification breast tumors as either malignant or benign based on medical data provided by the radiologist. The research aims to outperform the current predictive accuracy of the radiologist's assessments, providing a more reliable and faster indication of malignancy.

Unlike former studies, which mainly focused on training neural networks on raw mammogram images to classify the pathology of the tumors, our project mainly uses the radiologists' tabular data, which strictly contains information about the patients' breast density and tumor disregarding the patients' background.

2 Related Work

Breast cancer classification is a critical and broad area of research in both machine and deep learning. Throughout the years, various studies have been conducted to improve the accuracy of breast cancer diagnosis, leveraging different datasets, preprocessing techniques, and machine learning models.

2.1 DDSM Dataset

Several studies have used the DDSM dataset for breast cancer diagnosis, where most have focused on image-based classification rather than leveraging the structured data from radiologists' reports.

A 2019 research project titled *"Deep Learning to Improve Breast Cancer Detection on Screening* Mammography" (Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., & Sieh, W. ,2019) explored the use of neural networks, leveraging pretrained models such as VGG16 and ResNet50. The study demonstrated that incorporating both MLO and CC mammogram views significantly improved the AUC score, highlighting the importance of multiple perspectives in mammographic analysis. One of the study's limitations was the constraint posed by GPU capacity, which required downsizing the images to fit the available hardware. Our research aims to overcome this constraint by focusing primarily on tabular data instead of relying solely on mammogram images.

2.2 Breast Cancer Classification Studies

Many studies have been conducted on several Breast Cancer Datasets, some include the Wisconsin Breast Cancer Diagnostic (WBCD) dataset which features some of the tumors' characteristics such as area, perimeter, smoothness etc. The 'Machine Learning Based Comparative Analysis for Breast Cancer Prediction' (Chen, H., Wang, N., Du, X., Mei, K., Zhou, Y., & Cai, G. ,2023) study compares between the random forest, decision tree, K-nearest neighbor and logistic regression algorithms and reaching up to 98.6% accuracy with logistic regression.

2.3 Breast Cancer Factors

According to the *National Breast Cancer* Foundation (D. Shockney, L.,2024), two out of three women diagnosed with breast cancer are over the age of 55, underscoring the significance of age—an attribute missing from the DDSM dataset. Other key factors influencing the likelihood of malignancy include family history (whether close relatives have had breast cancer) and personal history, as individuals previously diagnosed with breast cancer are at higher risk of recurrence. Additional lifestyle factors, such as diet, daily physical activity, and body weight, also play a role in determining cancer risk.

3 Approach

3.1 Data Exploration

An analysis of the tumor labels (pathology diagnosis) has been conducted. The initial findings indicate a balanced distribution in the mass dataset, with 46.2% of tumors classified as benign and 53.8% as malignant. In contrast, the calcification dataset exhibits an imbalanced distribution, with 64.05% of tumors identified as benign (Figure 3.1). Additionally, the radiologist assessment scores are not distributed linearly in both datasets, as shown in Figures 3.2 and 3.3. Notably, the probabilities of a tumor being classified as benign or malignant in assessment 4 are nearly equal. Since this assessment accounts for just over 40% of the dataset, there is a critical need for enhanced classification and prediction methods for tumors. This indicates that more than 40% of the women lack clarity regarding their predicted pathology.

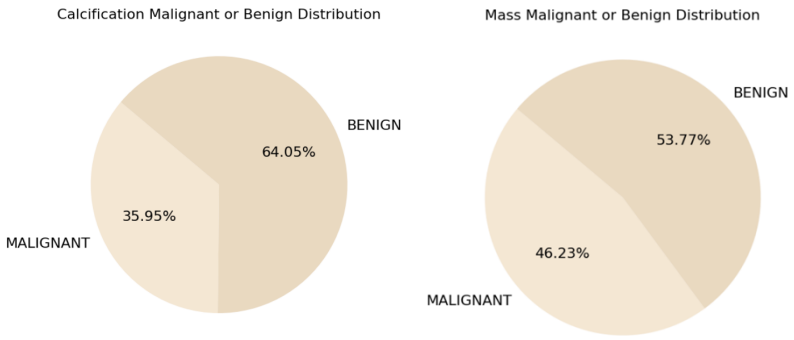


Figure 3.1.1 Pathology Labeling Distribution

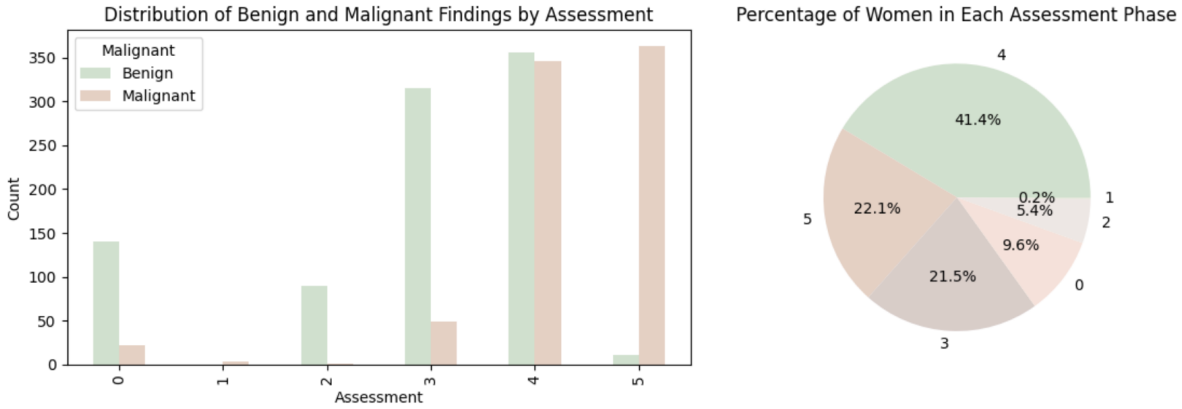


Figure 3.1.2 Mass Assessment Distribution

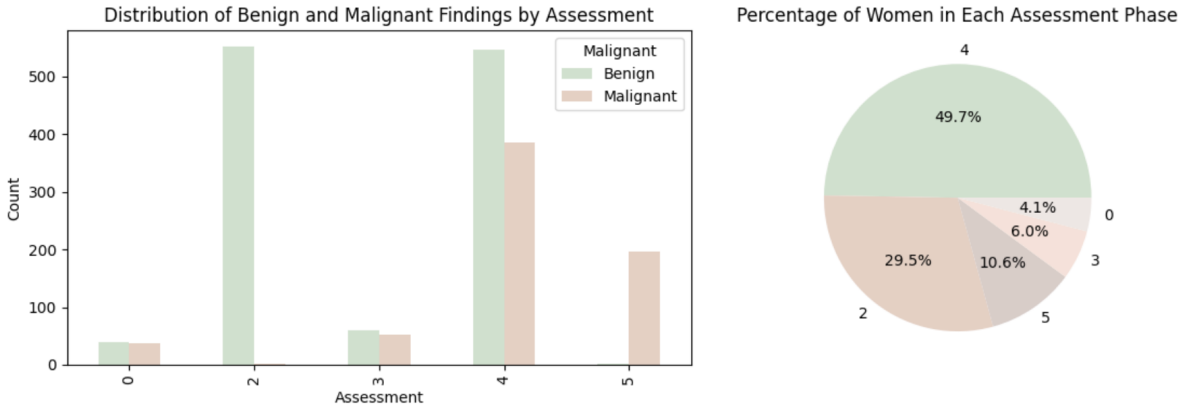


Figure 3.1.3 Calcification Assessment Distribution

Recognizing the significance of effective preprocessing, we identified and addressed several critical issues prior to model development. One of the initial challenges arose during our experimentation with the PyCaret library, which is generally quite beneficial for data analysis and research. The PyCaret's built-in function for model comparison produced accuracy scores around 90% for most of the models tested. However, upon further investigation, feature importance analysis (Appendix A) indicated that the patient ID column inappropriately inflated the accuracy, as the experiments pre-split the data into training and testing sets, allowing the possibility of tumors from the same patient appearing in both sets.

This scenario negatively impacted the reliability of the experiment. To address this issue, we removed the patient ID column and created a unique tumor ID for each case resulting in highest model accuracy of 78.7%. This change underscored the extent of the initial problem. Recognizing the limitations of PyCaret for our task, it was decided to proceed with a custom algorithm for model comparison.

3.2 Feature Engineering and Preprocessing Steps

Machine learning algorithms typically process data on a row-by-row basis and do not establish connections between individual rows. For instance, the ‘abnormality ID’ column, which contains integers ranging from 1 to 6 that denote the tumor number in a patient's breast, is not inherently understood by the algorithms as it is by human interpreters. Specifically, if a patient has a row with an abnormality ID of 4, it indicates the presence of four tumors in that breast. To effectively utilize this information and potentially enhance model performance, we introduced a binary column to indicate whether each patient had more than one tumor. Furthermore, some patients had tumors classified by radiologists using multiple shape descriptors, represented as hyphen-separated categories (e.g., "ROUND-IRREGULAR-ARCHITECTURAL_DISTORTION"). We addressed this by splitting these multi-shape labels into separate categorical features using one-hot encoding.

Next, one-hot encoding was applied to convert the remaining categorical variables into numerical features. This process also included the breast density column which contained numbers between 1 and 4, ensuring consistency with the binary nature of other columns. Finally, the Assessment and Subtlety columns were excluded from the training and testing dataset, as these represent the radiologists' assessments, which we did not want to bias the model with at this stage.

3.3 Model Training procedure

To compare model performance, we evaluated a range of classification algorithms: Logistic Regression, Linear Discriminant Analysis, Ridge Classifier, AdaBoost Classifier, Extra Trees Classifier, LightGBM, XGBoost, Random Forest, K-Neighbors Classifier, Decision Tree, Support Vector Machine (SVM) with a linear kernel, and Naive Bayes. These models were selected based on their inclusion in the model comparison function of the PyCaret library that was previously utilized.

We implemented a robust evaluation procedure by repeating the experiment 100 times and employing 10-fold grouped cross-validation. This approach allows us to utilize the entire dataset for both training and testing, ensuring sufficient data for analysis. The 10-fold split involved dividing the patient IDs into 10 groups and selecting all rows associated with these IDs for each specific fold. This method ensures that tumors from the same patient do not appear in both training and test sets, thus preventing data leakage. Additionally, shuffling the data enhances variability across folds. It is important to note that GroupKFold consistently generates the same splits, which diminishes the benefit of conducting 100 iterations; therefore, we implemented a manual splitting method.

For each fold, 9 groups were used for training, and the remaining group was used for testing. This process was repeated 10 times, with results stored for comparison. We performed this

100 times to reinforce the statistical reliability of our results. For each run, a table was created, saving the tumor ID, true labels, predicted labels, predicted probability and fold number for later analysis.

3.4 Hyperparameter Tuning

After selecting the top three models based on accuracy and AUC scores, we proceeded with hyperparameter tuning to further optimize model performance. The used method is GridSearchCV, splitting the data into 90% training and 10% test, similar to the earlier setup. For grid search cross-validation, we used a 9-fold CV to evaluate various combinations of hyperparameters on the training data, meaning the model is trained across 80% of the data, validated against the remaining 10% and finally tested. Grid search allowed us to systematically explore different parameter values, ensuring the best possible configuration for each model. The optimal parameters were then applied to the test data and the results were exported into CSVs.

4 Experiments

4.1 Datasets

The DDSM dataset was used in this project, consisting of two mass case files and two calcification (calc) case files, pre-split into training and test sets. The initial step involved merging the training and test data for both mass and calc cases to work with the full dataset. The mass dataset included 1,318 training cases and 378 test cases, which were combined for a comprehensive analysis. The calcification dataset includes 326 test cases and 1546 train cases which were also combined into one dataset.

4.2 Evaluation Method

When dealing with medical predictions, it is usually preferred to be on the safer side of the predictions and use the recall as the evaluation, meaning, doctors would usually rather misclassify a healthy woman with cancer and later in pathology to learn they were wrong, rather than telling a woman she's healthy and skip pathology, potentially letting the malignant tumor develop.
When the doctor suspects a woman has cancer, she'll definitely be sent to get the pathology done, the aim of this project is to try and classify as many tumors correctly as possible as the pathology results can take up to 14 days and we would like to avoid unnecessarily 'scaring' women during this period of time. Keeping this goal in mind, the evaluation method chosen is a combination of both the accuracy together with AUC score. It is important to note that AUC works well also for imbalance data, which is exactly the case in the calcification set, therefore using a combination of both scores returns a trustable estimation of the models.

4.3 Structured Data Model Training

The first step in the modeling training and testing process, was creating a custom compare_models() function to meet the specific needs of the study. Unlike the PyCaret's function, the custom function allows to divide the data into training and testing data as we needed in addition to saving the results of each model into a file for later analysis. The main

struggle of the project took place during this section as there are many premade functions, however each one led us to a different dead end missing a feature we couldn't compromise. After having decided to create a customize function, it was much easier later adding any feature needed.

After having run our function 100 times, resulting in each tumor being classified 100 times per model while having 12 models testing, we had sufficient data to compare and analyze. During the first step, the model trained on the following X-features entered by the radiologist: the mass shape and margins (after having been split one hotted), whether there's one or more tumors in the specified breast, the patient's breast density, the breast side in which the tumor is found and if the image is MLO or CC (note, it is important to keep both rows, as some tumors like that of patient 'P_01741' in the mass dataset could look round and spiculated in the CC mammogram but irregular and ill-defined in the MLO scan. The model trained on the malignant column which is a binary column giving 0 to all benign tumors and 1 for malignant.

4.4 Region of Interest (ROI) Images Feature extraction

A challenging yet teaching part of the project was aiming to extract features from the ROI images (see examples in Appendix B) already found in the dataset, the ROI images were supposed to be binary, however a few values other than 0 and 255 were spotted. Since the image dataset is heavy, the usage of Kaggle notebooks came in very handy, allowing us to run our code and extract the features without having to download them.
The features extracted were the area of the tumor, perimeter, longest and shortest axis. Additionally, circularity was computed as a value between 0 and 1, where 1 indicates a perfect circle, while eccentricity measures how elongated the shape is, with 0 representing a circular shape and 1 representing a line. The aspect ratio, defined as the ratio of the longest axis to the shortest axis, provides another measure of shape.
Other features included the convex hull area, which captures the area of the smallest convex shape enclosing the tumor, and solidity, the ratio of the tumor area to its convex hull area. Compactness quantifies the shape's regularity using its perimeter and area, and jaggedness assesses margin irregularity, with values close to 1 indicating irregular margins. For the formulas used to calculate these features, refer to Appendix E.

5 Results

5.1 BIRADS Tabular Data

As the project has been divided into 2 datasets, and we've been treating them separately throughout the entire project, we will also discuss the results separately starting with the Mass Dataset.
The use of structured, tabular data in this project not only allows for accurate classification but also enables greater interpretability as it offers insights into how the model arrives at the results. Such interpretability is often lost in image-based approaches, where the underlying features used by deep learning models can be difficult to understand.

After evaluating all models using the custom comparison function (detailed results are provided in Appendices C and D for the Mass and Calcification datasets, respectively), hyperparameter tuning was performed on the top three models: AdaBoost Classifier, Logistic Regression, and Linear Discriminant Analysis. Among these, AdaBoost achieved the highest performance, with an accuracy of 0.7999, a recall of 0.7928 (significantly improved from 0.7750 after tuning), and an AUC score of 0.8598. These results indicate that the model is stable and reliable—an essential quality when working with medical data and predictions. However, the Calcification dataset received lower scores, with AdaBoost remaining the top-performing model after hyperparameter tuning, resulting in an accuracy of 0.7275 and an AUC of 0.8045.

5.2 Results with Radiologist's Assessment

After having received our results, we were interested in finding how much not ignoring the 'Assessment' column in the training process would affect the final classification, we expected the accuracy to increase as having both the radiologist's assessment and the machine's classification could possibly be used together like ensembled learning which is usually known for well increasing the final AUC and accuracy. As expected, the AdaBoost model has managed to get a higher accuracy and correctly classified just over 83.1% of the entire data and the AUC score jumped to 0.9041 which is an incredible increase, also passing the 90% meaning the model is very reliable. The radiologist's tumor classification into assessments between 0 and 5 is indeed very meaningful as doctors may have seen more tumors than found in this current dataset and could find more patterns than the ones found by the machine, as well as the fact that the radiologists have more information which is lacking in the current dataset which is discussed in conclusion. The main disadvantage of using this column is that the final classification greatly depends on how good the radiologist is, meaning that the model cannot necessarily be used for women getting tested at a different clinic.

5.3 ROI Feature Extraction

The DDSM not only includes the radiologist's tabular data but also the mammogram images. Each tumor usually includes two mammograms, MLO where the scan is taken from the center of the chest outward, while the CC view is taken from above the breast (National Institute of Health – NIH, 2017). This paper uses the ROI images for feature extraction to have more information about the tumor and by that to possibly increase the accuracy. After having researched feature extraction in past written papers, the features this paper uses are area, perimeter, longest axis, shortest axis, circularity, eccentricity, aspect ratio, convex hull area, solidity, compactness and jaggedness (Appendix E for calculations).

The analysis began with the mass dataset, where using only the DDSM tabular data, we achieved an accuracy of 79.5% and an AUC score of 0.8624. After incorporating shape-related features, Logistic Regression emerged as the best-performing model, achieving an accuracy of 80.51% and an AUC of 0.86699 (see Appendix F for detailed results). Interestingly, some models, such as AdaBoost, experienced a decline in performance, with its AUC dropping to 0.845458. When running the shape features alone, without the DDSM descriptors, all models yielded accuracy scores below 60%, with Logistic Regression achieving the highest AUC of 0.624144 (Appendix G). As shown in Figure 5.3.1, the higher

correlations are found in the BI-RADS while the correlations with the features extracted from the images are lower.

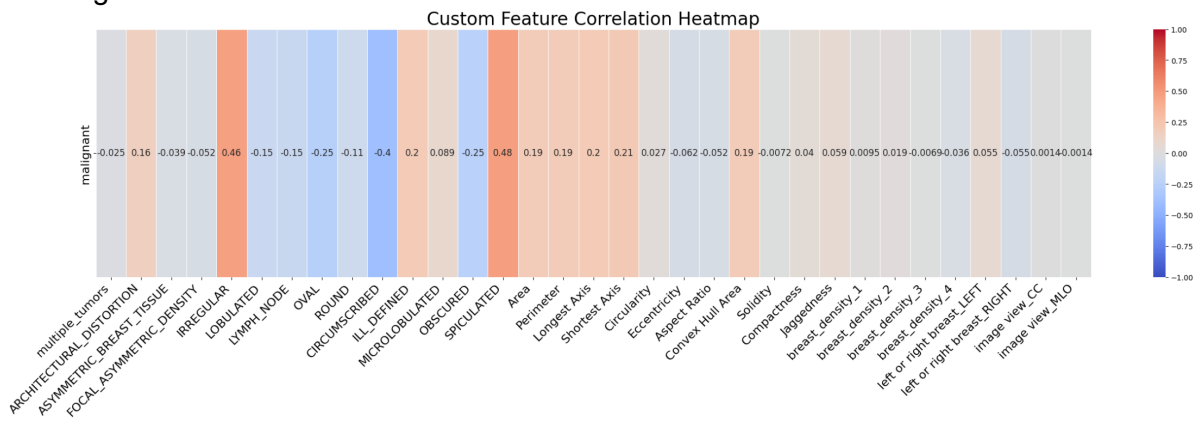


Figure 5.3.1 Mass Feature to Malignancy Correlation

Moving on to the calcification ROI images, the Kaggle data had missed just over 300 ROI images, therefore we've used a set of images from the original dataset. These images have been precheck on the mass data and managed to receive the same score give or take 1%. The other source had 30 images missing therefore we used the Kaggle images for mass, however, when dealing with the calcification data, Kaggle had many missing images and the original source only missed 66 images, therefore we choose this source.

In the calcification dataset, the features extracted from the ROI images alone achieved an accuracy of 73.22% and an AUC score of 0.782056 with the Linear Discriminant Analysis model (see Appendix H for all models), which outperformed the results obtained with the mass data. Notably, using only the BI-RADS features for the calcification dataset resulted in lower performance (Appendix G). However, combining both BI-RADS features, and ROI-extracted features led to AdaBoost becoming the top-performing model, achieving 77.23% accuracy and an AUC score of 0.8525, with improvements across all metrics (Appendix I). Figure 5.3.2 illustrates that the correlation between malignancy and the BI-RADS features is similar to the correlation between malignancy and the extracted features, which explains why their combination resulted in better accuracy and AUC than each one separately.

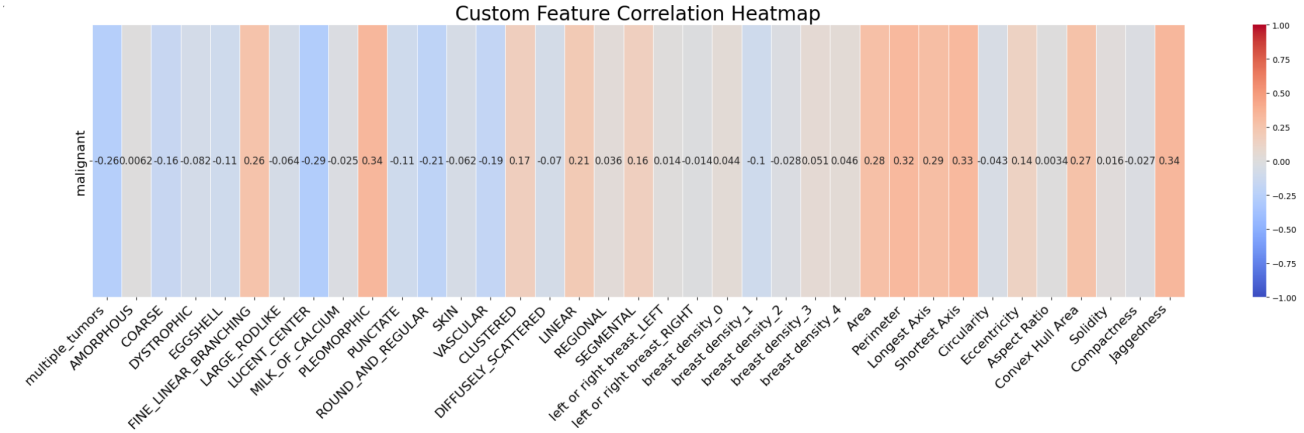


Figure 5.3.2 Calcification Feature to Malignancy Correlation

Given that the dataset is imbalanced, we needed to verify that the model's high scores were a result of effective learning rather than bias from class imbalance. To address this, we randomly down-sampled the benign class to match the size of the malignant class. After re-evaluating the models, both accuracy and AUC increased by nearly 1%, reaching 78.00% and 0.8677, respectively (see Appendix J for all models).

The results above arise the question of why the mass ROI features received such low scores? Could the data be faulty? Or rather the shape and other features extracted from the mass mammograms are less relevant to the malignancy of the tumor in compared to the correlation between the malignancy of calcification tumors and their shapes.

6 Conclusion and Further Work

In this study, we explored the DDSM dataset and applied machine learning models to classify breast tumors, achieving better results than initially anticipated. While the models did not achieve perfect accuracy—an expected outcome—this limitation can be attributed to missing key attributes in the dataset.

One critical missing feature is the patient's age, which would likely enhance the model's performance. As highlighted in related research, other factors such as family history and individual medical history also play crucial roles in breast cancer prediction but are not available in this dataset. Incorporating radiologists' assessments, which inherently account for such contextual information, could further improve classification results.

Recent studies have demonstrated the potential of neural networks for mammogram analysis, emphasizing the value of deep learning approaches. Combining our dataset with additional patient information, along with neural network-based models, could yield more accurate and comprehensive classification results.

A promising avenue for future work involves investigating the potential for benign tumors to become malignant over time. If predictive models could identify tumors at risk of becoming malignant, preventive interventions could significantly improve patient outcomes by enabling early treatment and reducing the likelihood of progression to cancer.

References

1. Awsaf. (n.d.). CBIS-DDSM: Breast Cancer Image Dataset. Www.kaggle.com. https://www.kaggle.com/datasets/awsaf49/cbis-ddsm-breast-cancer-image-dataset

2. Chen, H., Wang, N., Du, X., Mei, K., Zhou, Y., & Cai, G. (2023). Classification Prediction of Breast Cancer Based on Machine Learning. Computational Intelligence and Neuroscience, 2023, 1–9. https://doi.org/10.1155/2023/6530719

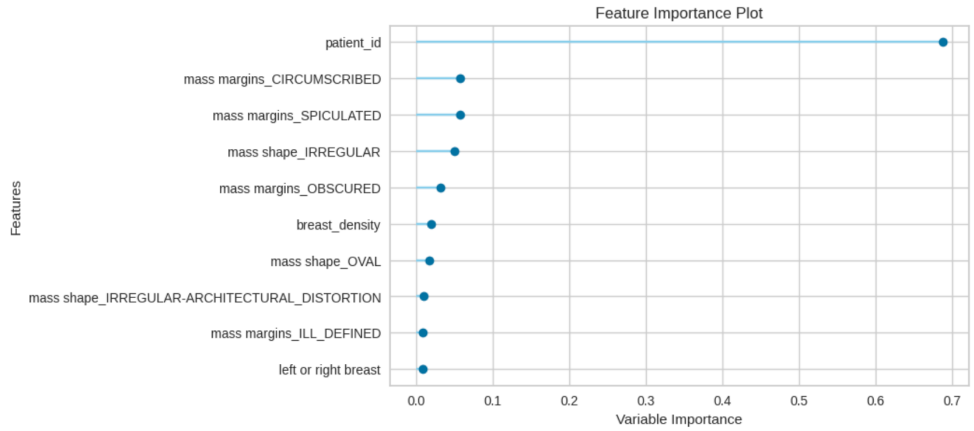
3. D. Shockney, L. (2024, October 8). Risk Factors. National Breast Cancer Foundation. https://www.nationalbreastcancer.org/breast-cancer-risk-factors/

4. GridSearchCV. (2024). Scikit-Learn. https://scikit-learn.org/dev/modules/generated/sklearn.model_selection.GridSearchCV.html

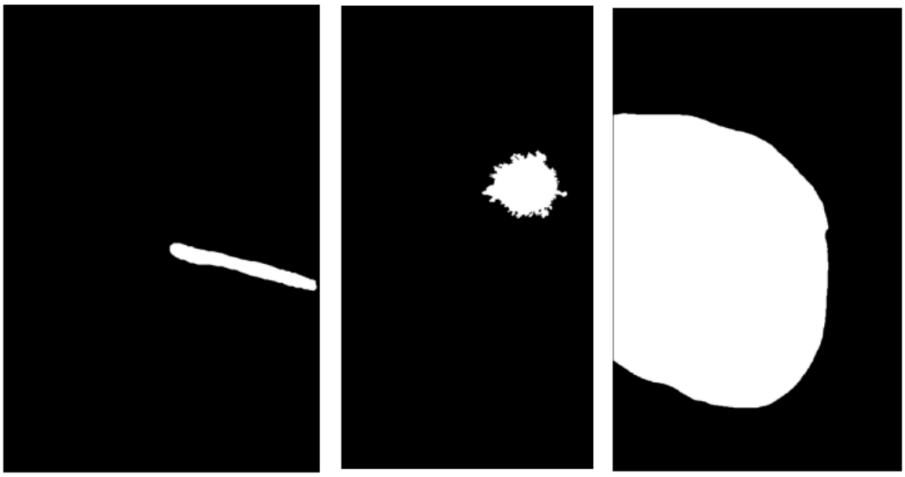
5. Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., & Sieh, W. (2019). Deep Learning to Improve Breast Cancer Detection on Screening Mammography. Scientific Reports, 9(1). https://doi.org/10.1038/s41598-019-48995-4

6. World Health Organization. (2024, March 13). Breast Cancer. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/breast-cancer

Appendix A: PyCaret Compare Models Feature Importance



Appendix B: ROI Images Examples



Appendix C: Mass Model Comparison Results

Model	Metric	accuracy	accuracy_std	auc	auc_std	f1_score	f1_score_std	precision	precision_std	recall	recall_std
Ada Boost Classifier		0.795000	0.004066	0.862486	0.001988	0.777545	0.004471	0.780091	0.005017	0.775038	0.005766
Decision Tree		0.771663	0.005786	0.804410	0.007726	0.744172	0.006781	0.771839	0.007221	0.718469	0.008743
Extra Trees Classifier		0.776669	0.004062	0.817603	0.005177	0.751091	0.004754	0.774660	0.005383	0.728954	0.006779
K Neighbors Classifier		0.773909	0.006604	0.828137	0.004388	0.750493	0.008570	0.766052	0.009316	0.735816	0.015944
LightGBM		0.786032	0.005071	0.853385	0.003823	0.768690	0.005451	0.768309	0.006570	0.769120	0.007291
Linear Discriminant Analysis		0.788650	0.003843	0.861053	0.001926	0.777068	0.004062	0.758275	0.004802	0.796849	0.005861
Logistic Regression		0.794298	0.003244	0.861859	0.001833	0.778000	0.003559	0.776293	0.004399	0.779745	0.005411
Naive Bayes		0.532889	0.005105	0.837273	0.002329	0.657084	0.002740	0.497321	0.002825	0.968099	0.003142
Random Forest		0.781067	0.004615	0.836246	0.003974	0.762583	0.004985	0.764586	0.005931	0.760625	0.006469
Ridge Classifier		0.788343	0.003620	NaN	NaN	0.776487	0.003825	0.758542	0.004611	0.795332	0.005622
SVM - Linear		0.789440	0.001647	0.781812	0.005733	0.795424	0.001490	0.721971	0.001880	0.885523	0.002122
XGBoost		0.781816	0.004770	0.842013	0.004257	0.764419	0.005307	0.763091	0.005719	0.765791	0.007280

Appendix D: Calcification Model Comparison Results

Metric	accuracy	accuracy_std	auc	auc_std	f1_score	f1_score_std	precision	precision_std	recall	recall_std
Model										
Ada Boost Classifier	0.716261	0.005275	0.800168	0.003466	0.570543	0.005695	0.626032	0.011244	0.524250	0.007762
Decision Tree	0.693654	0.007734	0.756045	0.008911	0.509949	0.009549	0.600510	0.016913	0.443343	0.010119
Extra Trees Classifier	0.705935	0.005847	0.779679	0.006142	0.528814	0.008653	0.623870	0.012717	0.459034	0.009881
K Neighbors Classifier	0.687815	0.009507	0.757313	0.007135	0.556960	0.018371	0.569332	0.016917	0.547088	0.038454
LightGBM	0.701058	0.006036	0.791080	0.005134	0.530271	0.008912	0.609507	0.012056	0.469376	0.009761
Linear Discriminant Analysis	0.720091	0.004219	0.799440	0.003047	0.567104	0.005900	0.638743	0.008823	0.510000	0.007390
Logistic Regression	0.719418	0.005171	0.802841	0.002775	0.571454	0.005776	0.633894	0.011122	0.520342	0.007016
Naive Bayes	0.607548	0.002917	0.710285	0.001532	0.644088	0.001761	0.477842	0.001870	0.987741	0.002052
Random Forest	0.704167	0.006542	0.791105	0.005096	0.543446	0.009977	0.610523	0.012524	0.489807	0.011901
Ridge Classifier	0.721453	0.004219	NaN	NaN	0.564752	0.006840	0.644397	0.008397	0.502704	0.008331
SVM - Linear	0.707372	0.010428	0.773774	0.006643	0.527785	0.010456	0.630635	0.028665	0.455082	0.020828
XGBoost	0.704963	0.006417	0.790279	0.005569	0.532493	0.010160	0.618841	0.012928	0.467429	0.011489

Appendix E: ROI Feature Extraction Formulas

$$\text{Area} = \text{Number of pixels where pixel value} > 0$$

$$Perimeter = \sum Boundary\ pixels\ of\ the\ largest\ contour$$

$$Eccentricity = 1 - \left(\frac{\frac{Major\ Axis}{2}}{\frac{Minor\ Axis}{2}}\right)^2$$

$$Aspect\ Ratio = \frac{Longest\ Axis}{Shortest\ Axis}$$

$$Solidity = \frac{Convex\ Hull\ Area}{Area}$$

$$Circularity = \frac{4\pi \times Area}{Perimeter^2}$$

$$Compactness = \frac{Perimeter^2}{4\pi \times Area}$$

$$Jaggedness = \frac{Original\ Perimeter}{Smoothed\ Perimeter}$$

Appendix F: Mass BI-RADS and ROI features Results

Metric / Model	accuracy	accuracy_std	auc	auc_std	f1_score	f1_score_std	precision	precision_std	recall	recall_std
Ada Boost Classifier	0.783791	0.004976	0.845458	0.003338	0.764899	0.005760	0.768976	0.005602	0.760906	0.008062
Decision Tree	0.719098	0.010275	0.717768	0.010294	0.697349	0.011181	0.694717	0.011956	0.700140	0.014304
Extra Trees Classifier	0.778149	0.004659	0.852637	0.003381	0.761682	0.005189	0.756504	0.005377	0.766964	0.007059
K Neighbors Classifier	0.777152	0.004633	0.830365	0.003497	0.754267	0.005429	0.769238	0.005320	0.739898	0.007255
LightGBM	0.789175	0.005488	0.852407	0.003426	0.770950	0.006144	0.774412	0.006819	0.767577	0.008546
Linear Discriminant Analysis	0.798833	0.003334	0.865086	0.001797	0.786767	0.003522	0.771336	0.004053	0.802844	0.004628
Logistic Regression	0.805094	0.003064	0.866995	0.001701	0.790212	0.003382	0.786381	0.003642	0.794094	0.004579
Naive Bayes	0.645702	0.005141	0.835908	0.001935	0.710294	0.003505	0.570986	0.003714	0.939541	0.004280
Random Forest	0.792612	0.004480	0.859149	0.002546	0.775146	0.005056	0.777006	0.005303	0.773329	0.007013
Ridge Classifier	0.800761	0.003536	NaN	NaN	0.788514	0.003694	0.774097	0.004333	0.803495	0.004563
XGBoost	0.786050	0.006040	0.849278	0.003598	0.767503	0.006969	0.771098	0.006921	0.764005	0.009937

Appendix G: Mass ROI Features Only Results

Metric / Model	accuracy	accuracy_std	auc	auc_std	f1_score	f1_score_std	precision	precision_std	recall	recall_std
Ada Boost Classifier	0.571604	0.007321	0.584732	0.007095	0.485987	0.009948	0.545656	0.010014	0.438189	0.011840
Decision Tree	0.529310	0.011357	0.526946	0.011349	0.493225	0.013123	0.491003	0.012176	0.495625	0.016580
Extra Trees Classifier	0.559982	0.006734	0.571179	0.005662	0.495000	0.008248	0.527211	0.008293	0.466556	0.009660
K Neighbors Classifier	0.556787	0.006738	0.565983	0.005877	0.504413	0.007700	0.522070	0.007863	0.487959	0.008975
LightGBM	0.569870	0.007688	0.587028	0.007401	0.514707	0.009748	0.537873	0.008959	0.493508	0.011490
Linear Discriminant Analysis	0.595932	0.003684	0.626754	0.002872	0.493438	0.005323	0.586757	0.005526	0.425753	0.005879
Logistic Regression	0.595377	0.003256	0.624144	0.003199	0.450413	0.004646	0.605217	0.006217	0.358686	0.004405
Naive Bayes	0.589888	0.002048	0.582178	0.005798	0.363481	0.003731	0.643251	0.005739	0.253316	0.003045
Random Forest	0.565896	0.007644	0.579203	0.006834	0.497002	0.010226	0.535128	0.009507	0.464031	0.012254
Ridge Classifier	0.593255	0.002808	NaN	NaN	0.460033	0.004317	0.595393	0.004888	0.374834	0.004443
SVM - Linear	0.585112	0.002760	0.622170	0.003365	0.347970	0.005418	0.636106	0.007815	0.239503	0.004366
XGBoost	0.560454	0.008539	0.576122	0.008302	0.506849	0.010244	0.526497	0.010028	0.488699	0.012222

Appendix H: Calcification ROI features Only Results

Metric Model	accuracy	accuracy_std	auc	auc_std	f1_score	f1_score_std	precision	precision_std	recall	recall_std
Ada Boost Classifier	0.732154	0.005397	0.770262	0.005468	0.602482	0.008866	0.652421	0.008927	0.559725	0.010998
Decision Tree	0.663544	0.008412	0.638273	0.009369	0.540745	0.012427	0.535451	0.011399	0.546260	0.015579
Extra Trees Classifier	0.722530	0.004401	0.775318	0.004312	0.588498	0.007637	0.636726	0.007111	0.547130	0.009962
K Neighbors Classifier	0.708378	0.003658	0.749090	0.003853	0.577071	0.005757	0.608714	0.005846	0.548595	0.007392
LightGBM	0.725626	0.005765	0.774104	0.004837	0.598508	0.009193	0.637676	0.009070	0.563939	0.011038
Linear Discriminant Analysis	0.732243	0.002081	0.782056	0.001342	0.573313	0.003724	0.679208	0.004134	0.496000	0.004341
Logistic Regression	0.703926	0.002252	0.769541	0.001386	0.496938	0.004472	0.647451	0.005096	0.403221	0.004680
Naive Bayes	0.712486	0.001718	0.744707	0.005552	0.450167	0.003993	0.734561	0.005474	0.324534	0.003589
Random Forest	0.729208	0.005193	0.778816	0.004349	0.595881	0.008388	0.649470	0.008655	0.550519	0.009857
Ridge Classifier	0.716600	0.002493	NaN	NaN	0.531609	0.004767	0.663547	0.005026	0.443450	0.005001
XGBoost	0.722115	0.006226	0.769099	0.005112	0.596057	0.009794	0.630364	0.009731	0.565374	0.011885

Appendix I: Calcification ROI features and BI-RADS Results – Original Imbalanced Data

Metric Model	accuracy	accuracy_std	auc	auc_std	f1_score	f1_score_std	precision	precision_std	recall	recall_std
Ada Boost Classifier	0.772337	0.006229	0.852575	0.004915	0.673908	0.009075	0.701273	0.009982	0.648672	0.010725
Decision Tree	0.725692	0.009383	0.705242	0.010529	0.625120	0.013775	0.619777	0.013213	0.630779	0.018584
Extra Trees Classifier	0.769291	0.005656	0.854569	0.003988	0.668511	0.008524	0.698019	0.009326	0.641496	0.011172
K Neighbors Classifier	0.663987	0.004830	0.655583	0.004582	0.480456	0.007172	0.546986	0.008698	0.428397	0.007385
LightGBM	0.770620	0.005983	0.852910	0.004034	0.676647	0.008850	0.692261	0.009053	0.661802	0.011307
Linear Discriminant Analysis	0.772713	0.003107	0.849210	0.002583	0.661437	0.004827	0.719348	0.005584	0.612183	0.005917
Logistic Regression	0.761240	0.003911	0.846086	0.002995	0.633496	0.007322	0.714566	0.007121	0.569038	0.010214
Naive Bayes	0.698040	0.001588	0.656065	0.015420	0.345119	0.004076	0.808535	0.008773	0.219389	0.002996
Random Forest	0.769081	0.005403	0.851267	0.003856	0.662758	0.008975	0.704547	0.008410	0.625740	0.011983
Ridge Classifier	0.765570	0.003351	NaN	NaN	0.648649	0.005090	0.710581	0.006087	0.596672	0.005759
XGBoost	0.771451	0.006651	0.852535	0.004454	0.679373	0.009975	0.691547	0.009831	0.667725	0.013164

Appendix J: Calcification ROI features and BI-RADS Results – Balanced Data

Metric Model	accuracy	accuracy_std	auc	auc_std	f1_score	f1_score_std	precision	precision_std	recall	recall_std
Ada Boost Classifier	0.780000	0.005298	0.867722	0.003790	0.783134	0.005598	0.772113	0.004946	0.794504	0.007880
Decision Tree	0.728092	0.005492	0.728092	0.005492	0.727804	0.005142	0.728630	0.006876	0.727023	0.006137
Extra Trees Classifier	0.769618	0.006485	0.860853	0.003972	0.774068	0.006294	0.759423	0.006813	0.789313	0.007161
K Neighbors Classifier	0.619237	0.001013	0.656857	0.002839	0.612965	0.002195	0.623230	0.001164	0.603053	0.004529
LightGBM	0.765191	0.004866	0.858069	0.002557	0.769371	0.006398	0.755806	0.001694	0.783511	0.012037
Linear Discriminant Analysis	0.766565	0.003461	0.855574	0.001890	0.780847	0.003416	0.735826	0.003234	0.831756	0.005324
Logistic Regression	0.768550	0.006209	0.858310	0.002913	0.784654	0.005913	0.733612	0.005671	0.843359	0.007997
Naive Bayes	0.612366	0.001630	0.674379	0.008942	0.402158	0.003681	0.878602	0.003879	0.260763	0.002960
Random Forest	0.779695	0.004961	0.862524	0.002302	0.786606	0.005755	0.762605	0.002963	0.812214	0.009988
Ridge Classifier	0.758473	0.005885	NaN	NaN	0.777366	0.005619	0.720954	0.004702	0.843359	0.006869
XGBoost	0.764885	0.007063	0.856050	0.003847	0.767788	0.007065	0.758426	0.007046	0.777405	0.008102

Appendix K: Calcification Data BI-RADS only Including Tumors that have ROI Images

Metric Model	accuracy	accuracy_std	auc	auc_std	f1_score	f1_score_std	precision	precision_std	recall	recall_std
Ada Boost Classifier	0.706811	0.006551	0.797044	0.003532	0.571157	0.006210	0.608638	0.013156	0.538321	0.010145
Decision Tree	0.688583	0.006579	0.753800	0.009435	0.504506	0.009986	0.596727	0.014335	0.437221	0.012515
Extra Trees Classifier	0.698217	0.005069	0.776897	0.005729	0.523349	0.007292	0.612802	0.011349	0.456855	0.009745
K Neighbors Classifier	0.688583	0.008852	0.763924	0.006432	0.568877	0.019271	0.571896	0.015281	0.568031	0.041146
LightGBM	0.693798	0.006757	0.788109	0.005952	0.520917	0.009266	0.602406	0.013858	0.459023	0.010571
Linear Discriminant Analysis	0.710354	0.006338	0.795178	0.003401	0.567043	0.008894	0.619395	0.011893	0.523023	0.011356
Logistic Regression	0.707940	0.006135	0.798939	0.002752	0.577533	0.008882	0.607625	0.010941	0.550534	0.013645
Naive Bayes	0.606733	0.003026	0.707744	0.001994	0.645656	0.001896	0.479538	0.001942	0.987878	0.002273
Random Forest	0.697730	0.006395	0.789268	0.005190	0.540508	0.008721	0.602560	0.012559	0.490229	0.010955
Ridge Classifier	0.713610	0.005203	NaN	NaN	0.555686	0.008394	0.635526	0.011337	0.493893	0.012184
SVM - Linear	0.689502	0.010398	0.782835	0.006877	0.557263	0.014650	0.578256	0.020320	0.539573	0.032549
XGBoost	0.700321	0.006786	0.786471	0.005478	0.523290	0.009911	0.618713	0.014643	0.453557	0.011346