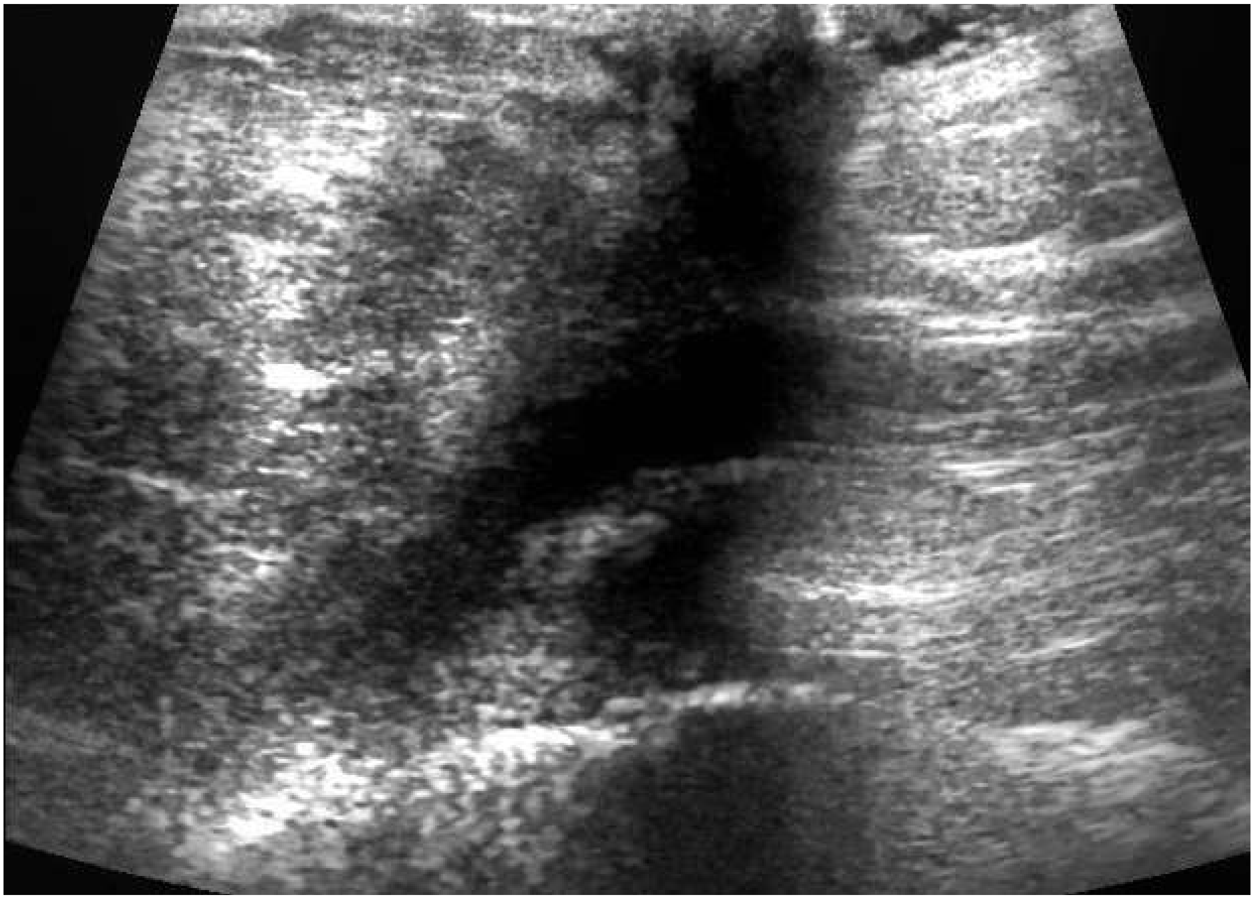
DATS 6303

Breast Cancer Ultrasound Classification

GROUP #5

Jack McMorrow, Alejandra Mejia



Introduction

Breast cancer remains a global health challenge, necessitating effective early detection strategies for improved patient outcomes. As the most prevalent cancer among women worldwide, the significance of early detection cannot be overstated. This study delves into the realm of breast cancer diagnosis, specifically leveraging deep learning approaches in ultrasound image analysis. The use of convolutional neural networks (CNNs), such as ResNet-50 and ResNet-101, showcases the transformative potential of data-driven and adaptive models in automating segmentation tasks. The study explores the dataset's origin, detailing the preprocessing steps undertaken for quality enhancement and compatibility with deep learning architectures. Additionally, it sheds light on the model selection process, training algorithms, and the role of data augmentation techniques and finally provides results of the approaches followed.

Literature Review

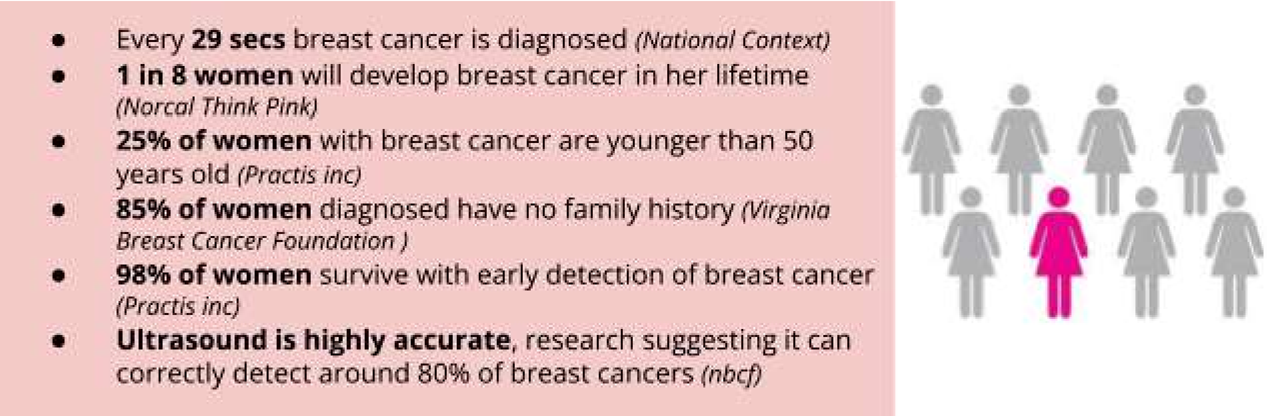
Breast cancer is a significant health concern worldwide, and early detection is crucial for improving patient outcomes. According to the World Health Organization (WHO), breast cancer is the most common cancer among women worldwide, with approximately 2.3 million new cases diagnosed in 2020 alone. Breast cancer predominantly affects women, but it can also occur in men, albeit at much lower rates

Early detection plays a pivotal role in improving breast cancer outcomes. According to the American Cancer Society, the five-year survival rate for localized breast cancer is around 99%, underscoring the significance of identifying tumors in their early stages. However, late-stage diagnoses remain a challenge, leading to lower survival rates and more aggressive treatment regimens.

While mammography has been a traditional method, ultrasound has emerged as a valuable adjunct in breast cancer detection. Ultrasound demonstrates superior sensitivity in detecting breast lesions, especially in women with dense breast tissue where mammography may be less effective (Berg et al., 2008). Unlike mammography, ultrasound does not involve ionizing radiation, minimizing potential risks associated with repeated exposure and making it a safer option for certain populations, including pregnant women

(Smith-Bindman et al., 2011). Moreover, ultrasound allows for real-time imaging and dynamic characterization of breast lesions, providing additional information on lesion vascularity, size, and morphology (Corsetti et al., 2008).

Figure 1. Quick glance at breast cancer stats



- Every **29 secs** breast cancer is diagnosed *(National Context)*
- **1 in 8 women** will develop breast cancer in her lifetime *(Norcal Think Pink)*
- **25% of women** with breast cancer are younger than 50 years old *(Practis inc)*
- **85% of women** diagnosed have no family history *(Virginia Breast Cancer Foundation)*
- **98% of women** survive with early detection of breast cancer *(Practis inc)*
- **Ultrasound is highly accurate**, research suggesting it can correctly detect around 80% of breast cancers *(nbcf)*

Manual interpretation of ultrasound images by trained radiologists can be a time-intensive process, requiring a thorough examination of images for the identification and characterization of potential abnormalities. Deep learning approaches offer significant advantages particularly in segmentation tasks. Deep learning models, although initially may demand time for training and validation offer the opportunity to automate segmentation. For example, models such as convolutional neural networks (CNNs), can automatically learn hierarchical representations of data directly from raw input. This ability to automatically extract relevant features from complex ultrasound images is crucial in capturing nuanced patterns associated with breast cancer lesions. The ability of deep learning models to automatically learn and adapt to variations in ultrasound images makes them well-suited for the inherent complexity of medical image analysis tasks.

In the context of breast cancer diagnosis and segmentation, the automated and adaptive nature of deep learning models not only enhances accuracy but also facilitates efficiency in healthcare workflows. By automating segmentation tasks, these models enable healthcare professionals to focus on more complex aspects of diagnosis and treatment planning, thereby contributing to expedited decision-making and improved patient outcomes (Ronneberger et al., 2015). Overall, the data-driven and adaptive nature of deep learning approaches positions them as powerful tools in advancing the field of breast cancer ultrasound image analysis.

Description of the Dataset

The dataset was found on Kaggle, which originated from a 2019 study published in India (Al-Dhabyani et al., 2019). The dataset generated from this study comprised breast ultrasound images collected at baseline from women aged 25 to 75 years old in 2018. The dataset included a total of 600 female patients, with 780 ultrasound images captured at a resolution of 500 × 500 pixels and stored in PNG format. The images are categorized into three classes: normal, benign, and malignant. Additionally, the authors made some preprocessing like fast photo crop to crop all images to different sizes, eliminating unused boundaries. Radiologists at Baheya hospital then reviewed and checked all images for quality assurance. Ground truth, in the form of image boundaries, was established for each image using Matlab and called a mask in the database. Figure 2, shows the distribution of the data raw dataset, while figure 3 provides a snapshot of the images available in the dataset.

Figure 2. Raw dataset class distribution

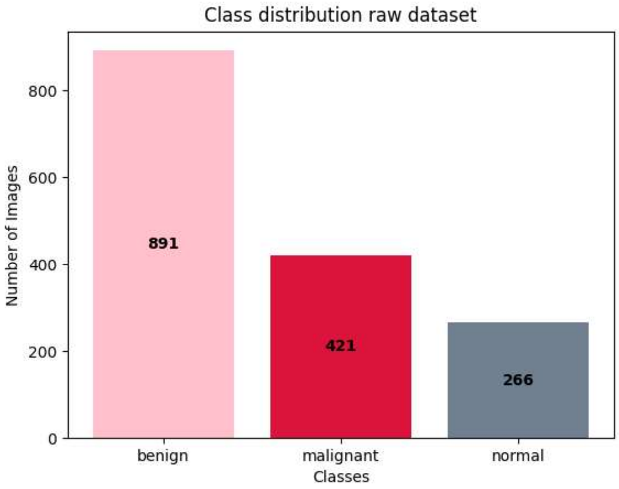
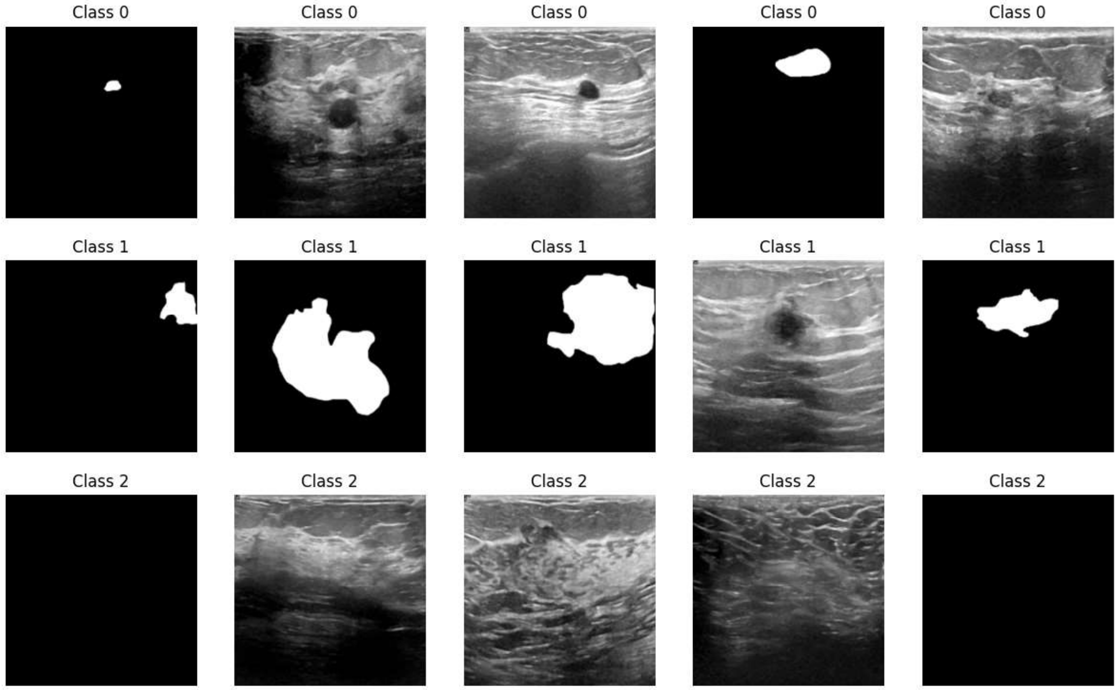


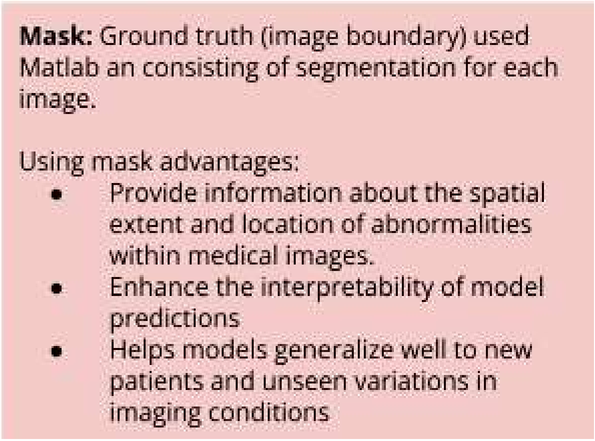
Figure 3. Raw database snapshot



Data Pre-processing

Preprocessing of the breast ultrasound dataset involved a series of tasks aimed at enhancing its quality, facilitating subsequent analysis, and ensuring compatibility with deep learning models. The initial step in the preprocessing pipeline involved the separation of the dataset into three distinct categories: original images, masked images, and a new dataset created by overlaying the original and masked images. This separation allowed for a nuanced exploration of both the raw data and the ground truth annotations provided by the masks (See Figure 4), essential for training and validating deep learning models.

Figure 4. What are masks and why are they important?



Mask: Ground truth (image boundary) used Matlab an consisting of segmentation for each image.

Using mask advantages:

- Provide information about the spatial extent and location of abnormalities within medical images.
- Enhance the interpretability of model predictions
- Helps models generalize well to new patients and unseen variations in imaging conditions

After the creation and separation of these datasets and exploration of the data was performed to gain insights regarding the new class distribution. Additionally, the sample images from each class with the overlaid mask were scrutinized to visually assess the diversity and complexity of the dataset. (See Figure 5 and 6)

Figure 5. Overlaid dataset class distribution

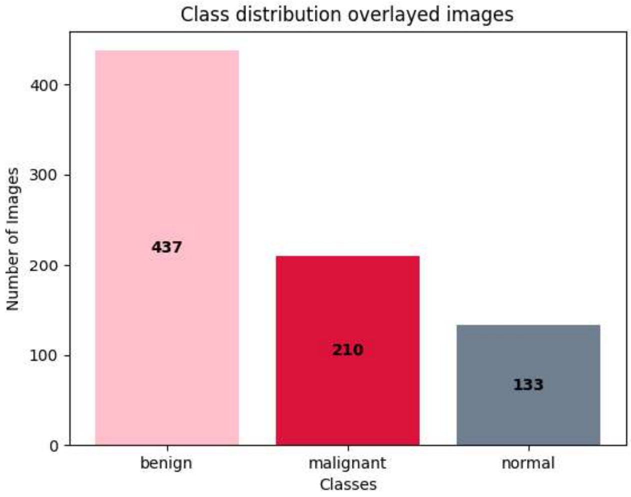
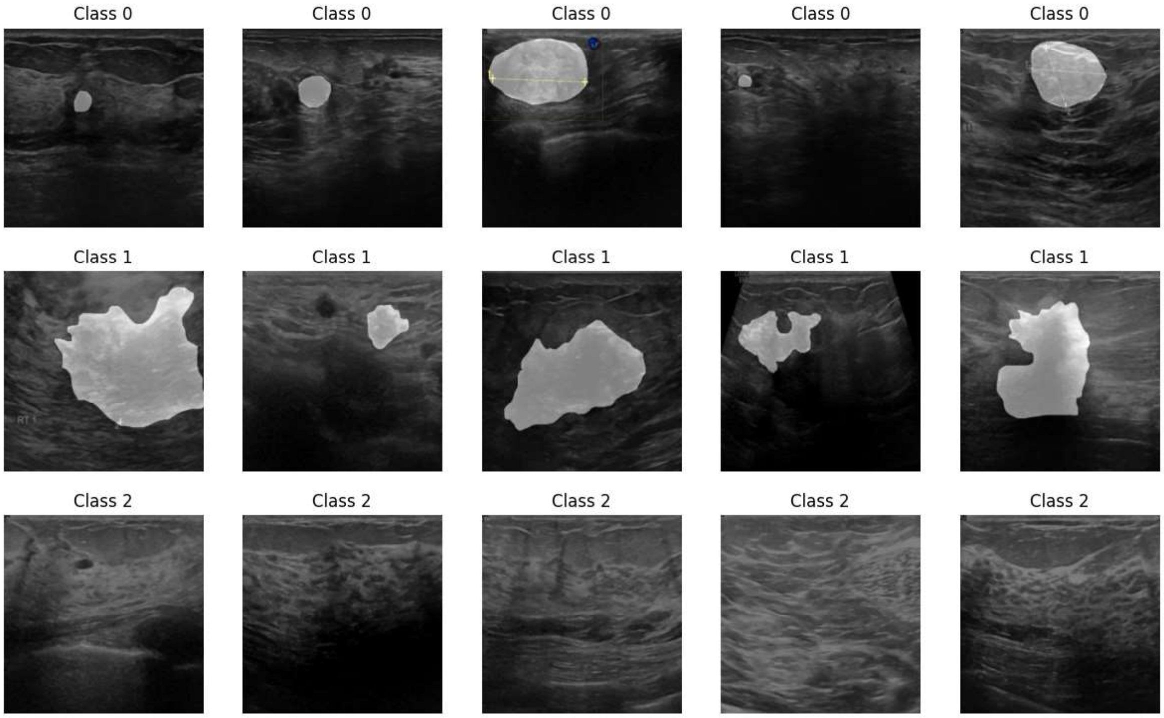


Figure 6. Overlaid dataset snapshot



Resizing of images to a standardized dimension of 224x224 pixels was performed as part of the preprocessing pipeline. This standardization aligned the dataset with the input requirements of deep learning architectures, specifically ResNet, facilitating seamless integration into the model training process.

Further exploration involved computing the standard deviation and mean values across the dataset. This information was crucial for normalization, a common practice in deep learning that ensures consistent and stable convergence during model training. The normalization process contributes to the model's ability to generalize effectively across diverse images. The normalization values that we calculated were 0.3301 for mean and 0.1993 for standard deviation.

In addition to the dimension-related preprocessing steps, labels were reassigned to streamline the classification task. The label transformation replaced the original classification (normal, benign, malignant) with a simplified numerical representation: 0 for benign, 1 for malignant, and 2 for normal. This label encoding simplified the model's interpretation of the target classes, enhancing the efficiency of subsequent training and evaluation phases.

Finally, the dataset was partitioned into training, validation, and test sets to facilitate robust model evaluation. The dataset splitting ratios were defined as 0.5 for training, 0.2 for validation, and 0.3 for testing. This partitioning strategy ensured an adequate amount of data for model training while allocating sufficient samples for unbiased model assessment and validation.

Considerations for model selection

In recent years, PyTorch has gained immense popularity in the deep learning community, establishing itself as a powerful and flexible framework for building and training neural networks. Capitalizing on PyTorch's extensive capabilities, our project sought to enhance breast cancer classification in ultrasound images through a comprehensive exploration of two distinct approaches. The first approach involved exclusively utilizing the actual image files, disregarding the mask files traditionally used for ground truth annotations, in an effort to evaluate the model's ability to classify without the aid of segmentation information. The second approach adopted a more nuanced strategy, overlaying the original images with their corresponding masks and utilizing these composite images for classification. This approach aimed to leverage both the raw image data and the ground truth annotations to potentially improve model performance.

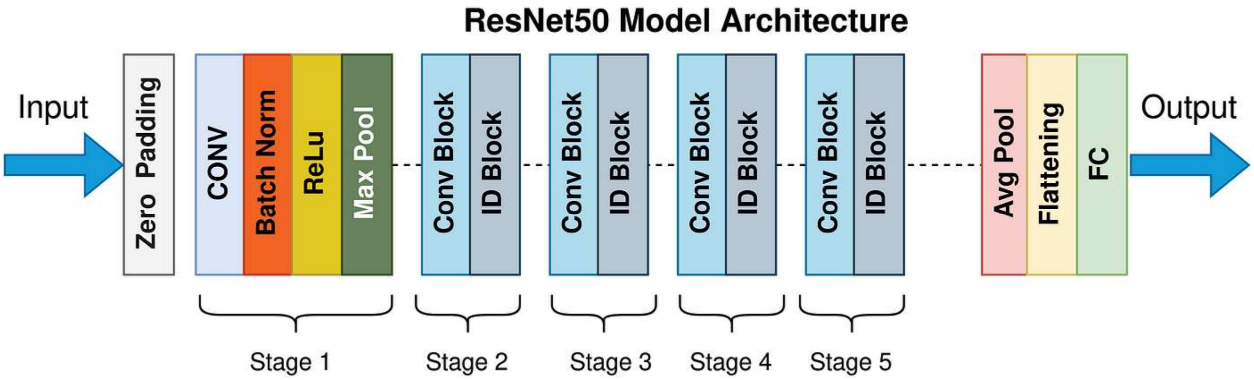
Training Algorithm

In the pursuit of optimized model architectures, we capitalized on the use of pretrained models, specifically ResNet-50 and ResNet-101, within the PyTorch framework. These pretrained models, trained on massive datasets, bring the advantage of already learned hierarchical representations of features, allowing for effective transfer learning. Leveraging the knowledge embedded in these pretrained models, our project aimed to harness the wealth of information they encapsulate to enhance the accuracy and efficiency of breast cancer classification in ultrasound images.

Resnet50

ResNet-50, a CNN architecture, has become a cornerstone in image classification tasks, including the analysis of breast ultrasound images in this specific study. Developed by Microsoft Research, ResNet-50 is a variant of the ResNet family known for its deep architecture, utilizing skip connections or residual blocks that enable the training of deep networks. In the context of image classification, ResNet-50 has demonstrated remarkable performance by addressing the challenge of vanishing gradients during training, a common issue in deep networks. With its 50-layer deep structure, ResNet-50 has the capacity to capture intricate hierarchical features and patterns in images, making it well-suited for discerning subtle differences in ultrasound textures indicative of normal, benign, and malignant cases. Leveraging residual learning, ResNet-50 facilitates the training of deeper networks, enhancing the model's ability to extract and learn complex representations, ultimately contributing to superior accuracy in image classification tasks, such as the discrimination of breast cancer lesions in ultrasound images

Figure 7. Resnet50 Architecture

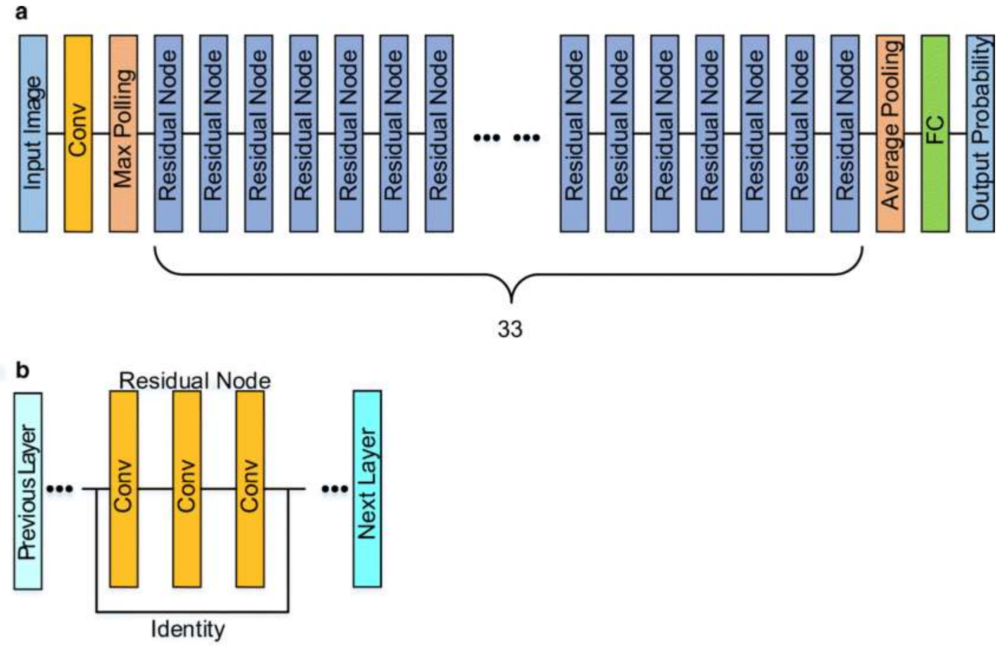


Source: Wikimedia commons

Resnet101

Developed as an extension of ResNet-50, ResNet-101 introduces a deeper network with 101 layers, allowing for an even more nuanced understanding of complex visual features. This extended depth enables the model to capture intricate hierarchical representations, making it particularly effective in discerning fine-grained details within images.

Figure 8. Resnet101 Architecture



Source:

Research Gate

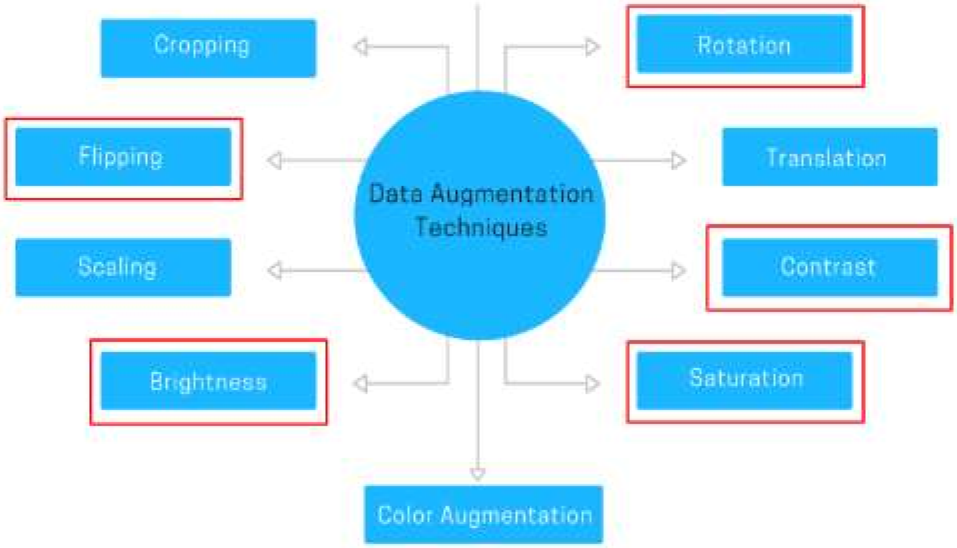
Model specification

The model architecture employed in the current project is based on the Adam optimizer, trained over 20 epochs using the CrossEntropyLoss function, and incorporates a learning rate adjustment strategy known as ReduceLROnPlateau. The Adam optimizer, known for its adaptive learning rates and efficient handling of sparse gradients, facilitates the optimization of model parameters. The choice of 20 epochs defines the number of times the entire dataset is passed through the model during training, striking a balance between achieving convergence and avoiding overfitting. The CrossEntropyLoss function is utilized as the optimization objective, aligning with classification tasks, where the goal is to minimize the difference between predicted and actual class probabilities. Additionally, the ReduceLROnPlateau strategy dynamically adjusts the learning rate during training, reducing it when the model performance plateaus. This adaptive learning rate strategy optimizes convergence, ensuring the model's ability to learn intricate patterns in the data efficiently. Collectively, these model specifications contribute to the overall efficiency and effectiveness of the breast cancer classification models in the project.

Data Augmentation techniques

The implementation of data augmentation techniques played a key role, and it became key when we implemented our deep learning models (ResNet-50 and ResNet-101). Traditional techniques include, cropping, rotation, flipping, translation, scaling, contrast, brightness, saturation and color augmentation. In the current project, where the objective is to classify breast cancer in ultrasound images, the diversity introduced by data augmentation was valuable. Five techniques were used. Flipping the images horizontally and vertically ensures that the model encounters a broad range of orientations, enhancing its adaptability to diverse ultrasound acquisition scenarios. Adjusting brightness, contrast, and saturation levels allows the model to discern features under varying imaging conditions, mirroring the complexities encountered in clinical settings. Rotation augmentation aids in simulating different probe positions during image acquisition. By employing these augmentation strategies, we not only augment the dataset but also enhance the models' capacity to learn intricate patterns and nuances, ultimately improving the accuracy and reliability of breast cancer classification with ResNet-50 and ResNet-101.

Figure 9. Data Augmentation Techniques



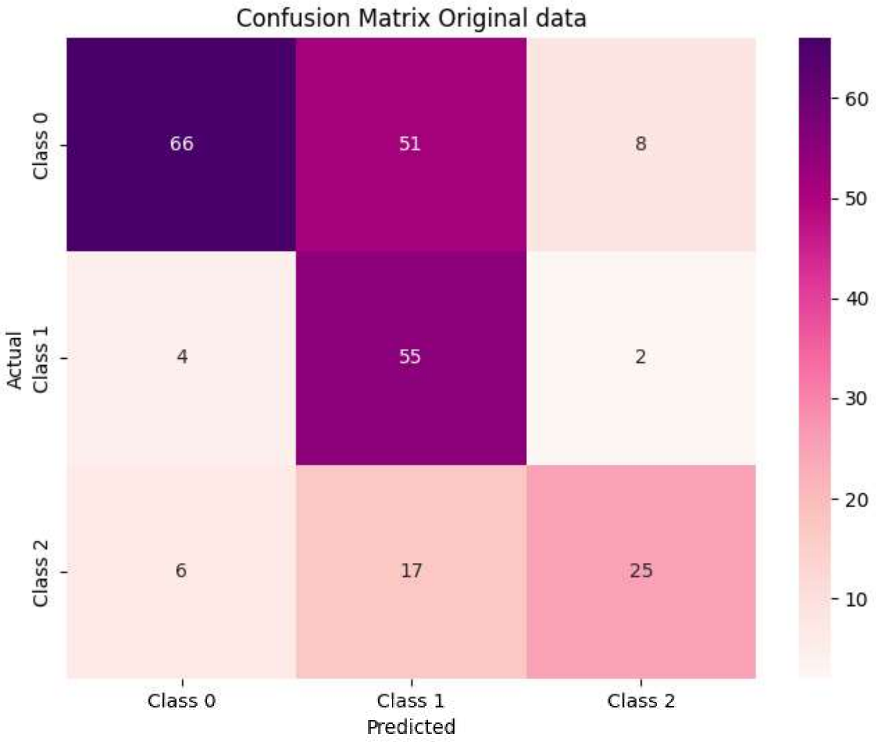
Source: AIMultiple

Results and Model Performance

Resnet50

Upon careful consideration of the outcomes derived from our initial models, we made a strategic decision to forego further exploration of ResNet-50. The confusion matrix for Model 1, while providing valuable insights, revealed moderate performance across the spectrum of breast cancer classes (normal, benign, malignant) (Figure 10).

Figure 10. Confusion Matrix for Model 1-ResNet 50



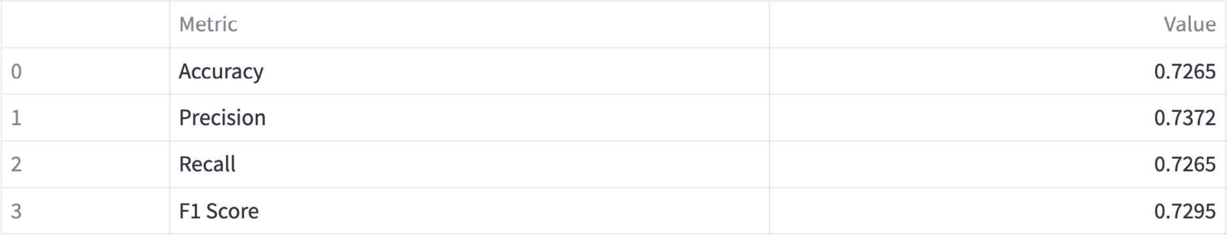
In contrast, the introduction of Model 2, incorporating overlays of original images with corresponding masks, presented a substantial improvement in classification accuracy, exemplified by its impressive confusion matrix (Figure 11). However, our results still did not reach an F1 greater than 90%, which is usually required for the expectiation of healthcare results.

Consequently, we redirected our efforts to explore the capabilities of ResNet-101, a more intricate architecture. Preliminary results indicate promising improvements over ResNet-50, motivating further investigation and potential adoption for enhanced breast cancer classification in ultrasound images.

Resnet101

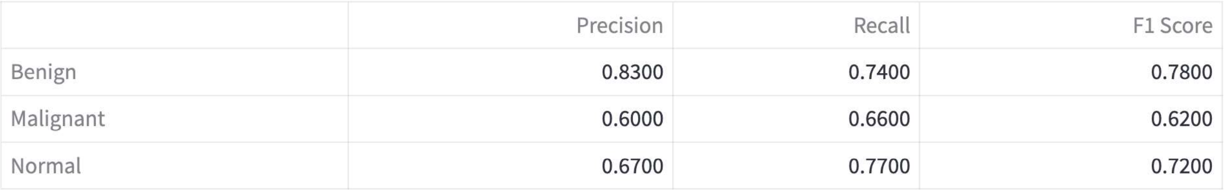
The original model using only the original ultrasound images proved to have moderate results. The macro F1 score for this model was around 0.73, with the classification performing best for benign images, followed by normal and then malignant. Complete results for this model are shown in Tables 1 and 2, with a confusion matrix in Figure 12.

Table 1. Macro Metrics for Performance of Model 1.



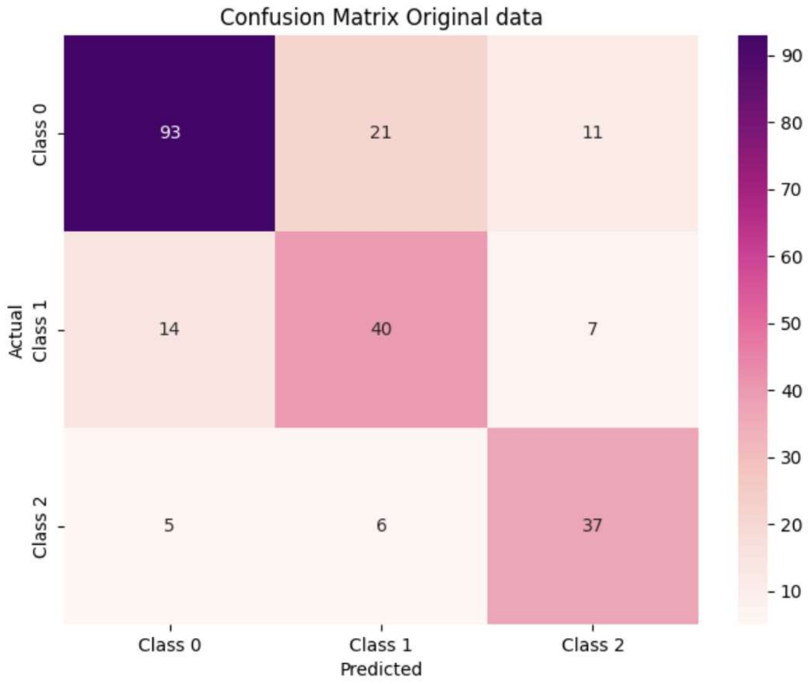
	Metric	Value
0	Accuracy	0.7265
1	Precision	0.7372
2	Recall	0.7265
3	F1 Score	0.7295

Table 2. Class Metrics for Performance of Model 1.



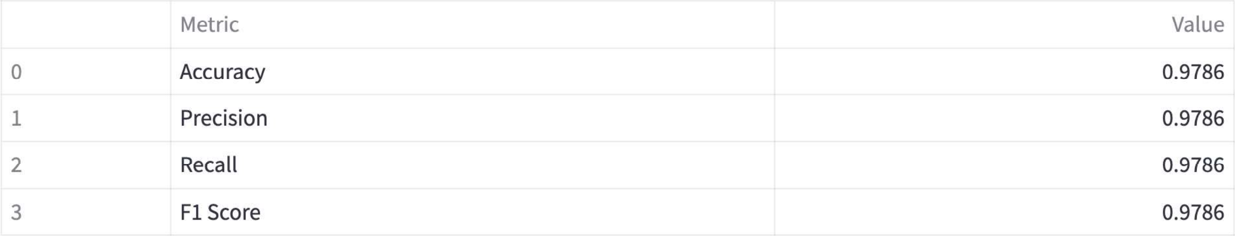
	Precision	Recall	F1 Score
Benign	0.8300	0.7400	0.7800
Malignant	0.6000	0.6600	0.6200
Normal	0.6700	0.7700	0.7200

Figure 12. Confusion Matrix for Model 1.



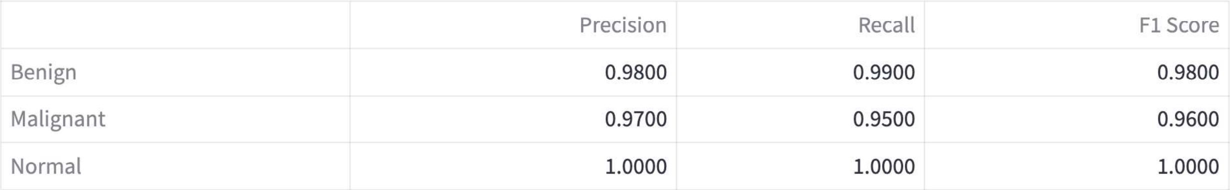
The second model we produced had much better results. This model had the same architecture as the original model, but we overlaid the original images with the corresponding mask images. This provided an image with a more defined outline of where the potential tumor would be. The macro F1 score for this model was 0.98, with good metrics performing across the board for all of the classes. Complete results are shown in Tables 3 and 4, with a confusion matrix in Figure 13.

Table 3. Macro Metrics for Performance of Model 2.



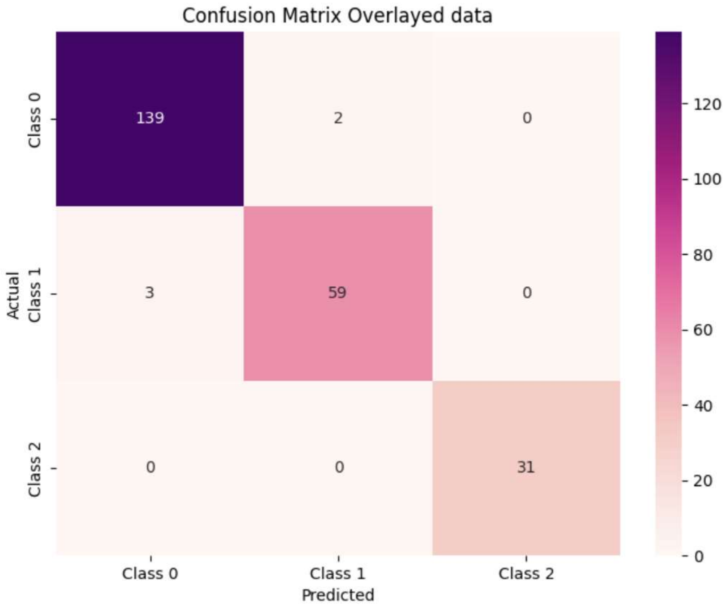
	Metric	Value
0	Accuracy	0.9786
1	Precision	0.9786
2	Recall	0.9786
3	F1 Score	0.9786

Table 4. Class Metrics for Performance of Model 2.



	Precision	Recall	F1 Score
Benign	0.9800	0.9900	0.9800
Malignant	0.9700	0.9500	0.9600
Normal	1.0000	1.0000	1.0000

Figure 13. Confusion Matrix for Model 2.



The training vs. validation losses for the two models are shown in Figures 14 and 15. The loss graph for the second dataset had inconsistent metrics across different epochs, which is likely due to the small size of the dataset that was used, which was a major limitation of our dataset. Furthermore, data augmentation, while enhancing the model's exposure to diverse patterns and improving generalization, might have introduced certain complexities that lead to non-uniform convergence. The diversity introduced through augmentation, such as random rotations, flips, and adjustments in brightness, may create instances where the model encounters variations challenging to reconcile within individual epochs.

Figure 14. Training vs. Validation Loss for Model 1-Resnet101

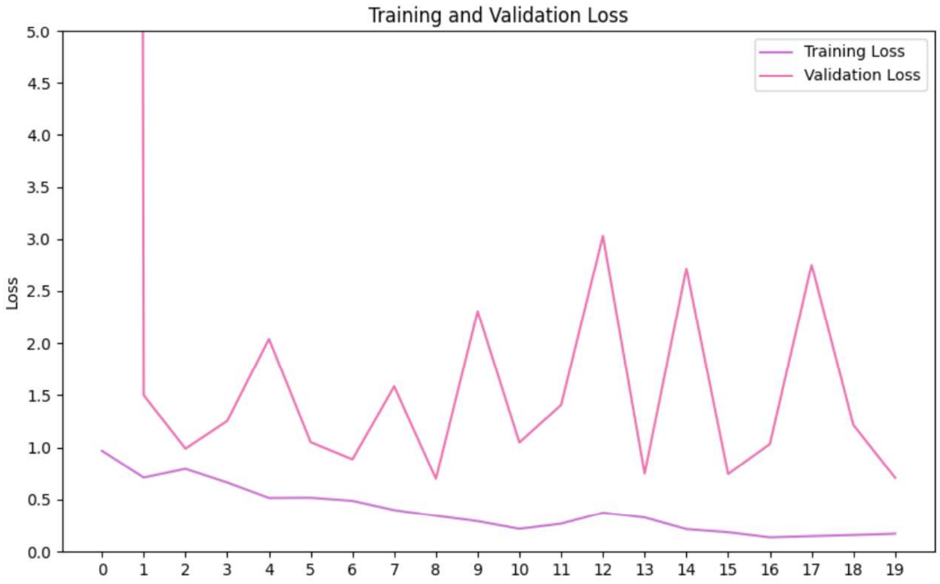
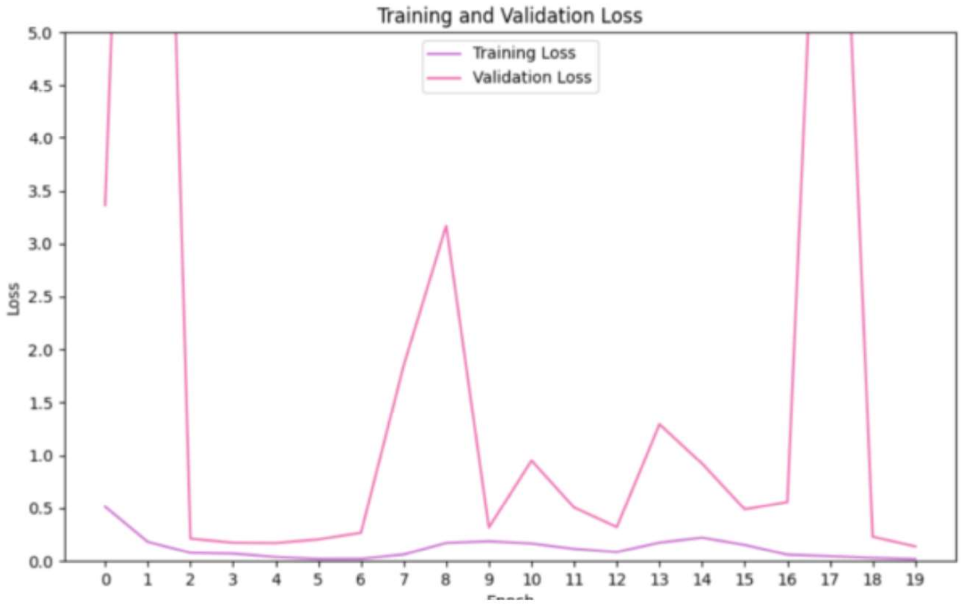


Figure 15. Training vs. Validation Loss for Model 2- ResNet101

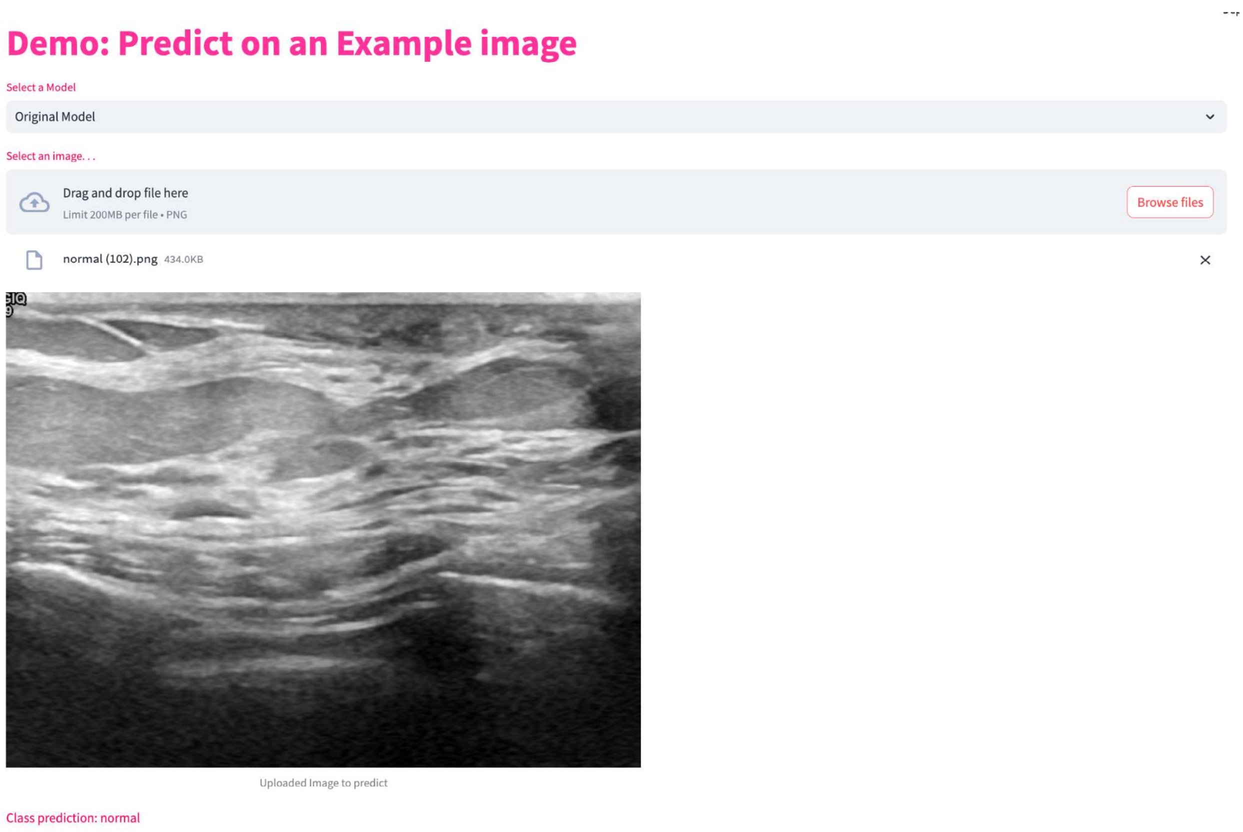


Overall, the second model performed very well and would likely be selected for further development if it were to be put into production for breast cancer screening methods.

Demo

As part of our presentation, we included a demo feature in the user interface that was developed in streamlit. Under the demo tab, an ultrasound image can be uploaded directly to the UI in .png format, and the model will run and create a prediction for you. This would allow a classification of benign, malignant, and normal to be easily made. This demonstrates the utility of using a model like this in a clinical setting, because once the image is produced a prediction could be made almost instantly.

Figure 16. Demo on UI to Predict Class of Image



Demo: Predict on an Example image

Select a Model

Original Model

Select an image. . .

normal (102).png 434.0KB

Uploaded Image to predict

Class prediction: normal

Summary and conclusions

Using the data provided to us, we were able to produce a model that has very good results for classifying breast cancer from ultrasound images. This demonstrates the utility of image classification systems to diagnose breast cancer. While current technologies are very good at classifying cancers, a technique like this allows for a method that is far less invasive than current cancer screenings. This could increase the availability of these screenings due to their convenience and low cost. This could dramatically increase the number of breast cancer cases detected. Early detection is key to properly treating breast cancer, so implementing techniques like this into a clinical setting could improve how we fight breast cancer.

Future work

Further refinement of the model and gaining deeper insights could have been achieved through the exploration of advanced techniques. Ensembling, a strategy involving the combination of predictions from multiple models, is a powerful approach that can significantly improve overall model performance and provide more robust predictions. Additionally, incorporating explainability techniques, such as SHAP (SHapley Additive exPlanations) or Grad-CAM (Gradient-weighted Class Activation Mapping), would have enhanced the interpretability of the model. These techniques contribute to understanding the model's decision-making process and identifying influential features in the input data. Evaluating the model's performance in real-world clinical settings is crucial to assess its practical utility, ensuring its seamless integration into existing healthcare infrastructure. This step is essential for validating the model's efficacy beyond controlled experimental conditions. Furthermore, considering ethical implications, addressing privacy concerns, and ensuring regulatory compliance in healthcare AI applications is imperative to establish responsible and ethical deployment practices. Taking these considerations into account would have added a comprehensive dimension to the overall project and its potential impact.

Limitations

The major limitation of our project was the size of our dataset. With only a few hundred images at our disposal, it makes it difficult to generalize the results of our model to the general population or to make it practical in a clinical setting. This is reflected in our training vs. validation loss graphs, which show a lot of irregularities in the validation data. Going forward, we would want to train these models on more data if they were to be used clinically.

Another limitation is the development of the ground truth images, which were essential to the training and predictions of our second model, which had much better results. The data source specifies the use of a matlab program to produce a segmentation of the tumors in the malignant and benign images. The intensity of this process is not described in the original source. If the masking process is rigorous, then the utility of our model might be decreased. If an expert is necessary to mask the images, then the model becomes less convenient. However, if these images are easy to produce, then it would be easy to overlay them over the original images and make an accurate prediction using the second model.

Further work could have improved the model and provided more insights. Techniques like ensembling, image generation, and more ensembling techniques could have improved the performance. We also wanted to experiment with segmentation techniques using tools like Grad-CAM. However, with only two group members and limited time we chose to focus on the main classification model as the core of this project.

References

American Cancer Society. (2022). "Breast Cancer Facts andFigures Report." Retrieved from https://www.cancer.org/research/cancer-facts-statistics/breast-cancer-facts-figures.html

Berg, W. A., Blume, J. D., Cormack, J. B., Mendelson, E. B., Lehrer, D., & Bohm-Velez, M. (2008). Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. JAMA, 299(18), 2151–2163.

Corsetti, V., Houssami, N., Ferrari, A., Ghirardi, M., Bellarosa, S., Angelini, O., ... & Remida, G. (2008). Breast screening with ultrasound in women with mammography-negative dense breasts: Evidence on incremental cancer detection and false positives, and associated cost. European Journal of Cancer, 44(4), 539-544.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer.

Smith-Bindman, R., Lipson, J., Marcus, R., Kim, K. P., Mahesh, M., Gould, R., & Miglioretti, D. L. (2011). Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. Archives of Internal Medicine, 171(22), 2071–2077.

World Health Organization. (2022). "Breast cancer: prevention and control." Retrieved from https://www.who.int/cancer/detection/breastcancer/en/

Appendix

Libraries:

Code File:

```
from Toolbox import DatasetCreator ,overlay_images_with_masks, CancerDataset, EDA,apply_augmentation_to_dataset, custom_collate_fn, predict
from sklearn.model_selection import train_test_split
from torch.utils.data import Dataset, DataLoader
from collections import Counter
import torchvision.models as models
import torch.nn as nn
import torch.optim as optim
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix, \
    classification_report
import torch
from tqdm import tqdm
from torch.optim.lr_scheduler import ReduceLROnPlateau
import random
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import confusion_matrix
from PIL import Image
```

```
import pandas as pd
```

Toolbox File:

```
import cv2
import numpy as np
from sklearn.model_selection import train_test_split
from torchvision import transforms
from torch.utils.data import Dataset, DataLoader
from PIL import Image
import tensorflow as tf
import torch
from collections import Counter
from torchvision.transforms.functional import resize
import random
import torchvision.models as models
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
import os
import shutil
from sklearn.model_selection import train_test_split
from torch.utils.data import Dataset, DataLoader
from collections import Counter
import torchvision.models as models
import torch.nn as nn
import torch.optim as optim
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix, \
    classification_report
import torch
from tqdm import tqdm
from torch.optim.lr_scheduler import ReduceLROnPlateau
```

Streamlite File:

```
import streamlit as st
import torch
import io
from PIL import Image
from Toolbox import predict
```