BIRADS Classification of Mammograms

Seshill Real Armas
Student# 1006961307
seshill.realarmas@mail.utoronto.ca

Shuhao Fan
Student# 1007998204
sh.fan@mail.utoronto.ca

Yazeed Bukhari
Student# 1007791579
yazeed.bukhari@mail.utoronto.ca

Ryan Benn
Student# 1007219225
ryan.benn@mail.utoronto.ca

Abstract

Breast cancer is one of the most prevalent forms of cancer in women worldwide, and the task of reading, detecting, and diagnosing malignant tumors is labor intensive. Recent deep-learning developments have helped ease the high labor demand for radiologists. We are aiming to build upon these models by pre-training with ResNet and GoogleNet, and then fine-tuning our model. —-Total Pages: 9

1 Introduction

Breast cancer is the most common - and second most lethal - cancer among Canadian women. Notably, in 2022, 25% of cancers diagnosed in women were breast cancers (Canadian Cancer Society, 2023b). This high incidence rate is alarming, and highlights the importance of establishing effective and efficient screening methods for the disease. Currently, screening mammograms (low-dose x-rays of the breasts) are used to check for cancer in individuals with no existing symptoms, and can help in identifying the disease in its early stages, at which point successful treatment of the cancer is more likely (Canadian Cancer Society, 2023a).

Given the high volume of mammograms conducted each year, it is important that opportunities for efficiency in the turn-around time of mammography results be explored; in particular, leveraging deep learning to improve the diagnosis process (U.S. Food and Drug Administration, 2023).

CADe (computer-aided detection) and CADx (computer-aided diagnosis) systems - jointly referred to simply as CAD systems - are programs that aim to assist radiologists in the detection and diagnosis of diseases, as well as in improving efficiencies in the screening process (Firmino et al., 2016). A type of key CAD system are those which independently learn the features they rely on to detect and diagnose abnormalities using deep learning (Guetari et al., 2023).

In regards to breast cancer, CADe has been used for several decades to assist in the detection of abnormalities in breast tissue. However, it has not been as successful as initially envisioned. In light of these short-comings, it is important to continue improving CAD systems.

The team plans to use convolutional neural networks to develop a model which will classify mammograms according to their Breast Imaging Reporting and Data System (BIRADS) score, where a score of one corresponding to negative findings and normal breast tissue, and a score of five indicating findings that highly suggest cancer (Daniel Liu, 2022). It is important to note that a BIRADS score of zero corresponds to cases where the mammograms results were inconclusive; as such, the team has chosen to focus solely on BIRADS scores of one through five.

2 Illustration

The project procedure is shown in Figure 1, and the data processing procedure is illustrated in Figure 2. Our baseline model is a normal CNN with 2 convolutional layers, 2 max pooling layers and 2 fully connected layers. The model is trained using ADAM optimizer to classify 5 classes of BIRADS. Our primary model combines 3 different deep learning models: Resnet50, GoogleNet and our baseline model. It is shown in Figure 3 and Figure 4 (reused in architecture).

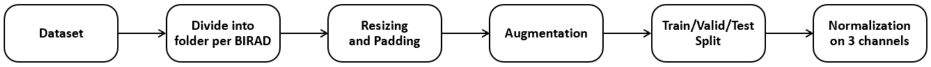


Figure 1: Data processing procedure

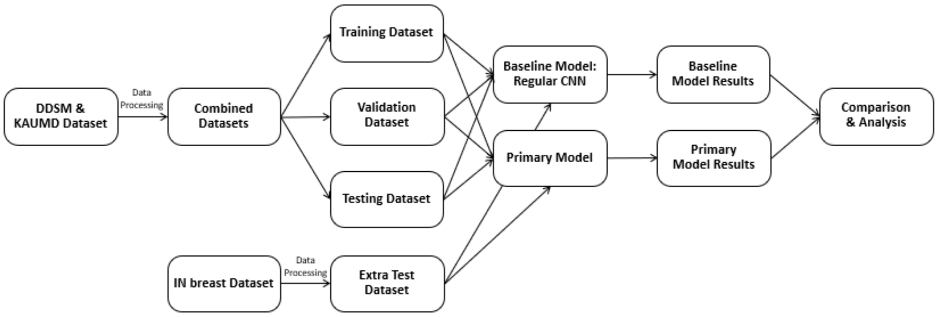


Figure 2: Project training procedure

3 Background & Related Work

Breast cancer is one of the most prevalent cancer types in the world, thus causing a high demand for radiologists who specialize in breast cancer imaging. To lower this flux of demand, computer-assisted models are being continually developed to increase diagnosis efficiency and accuracy. While these models can be extremely effective, they should not replace radiologists, but should instead be used as an accompanying tool.

Development started with traditional machine learning (ML) methods in the 1990s. This produced two major CAD model classes: CADe and CADx (Arka Bhowmik, 2022). CADe systems help identify the presence of malignant tumors, while CADx helps to classify and determine the diagnosis of the tumor (Ramadan, 2020). Unfortunately, as a result of the high false positive rate, these ML methods have since fallen out of favor with other deep learning (DL) models (Arka Bhowmik, 2022).

Following the breakthrough of AlexNet, deep neural networks have become the primary focus of development. A subset these neural networks, CNNs, are the primary DL method used for medical image classification as a result of their segmentation, feature recognition, and pattern identification.

Recently, in 2022, an accurate deep-learning model was developed using a CNN-CBR system to read and detect malignant tumors in mammograms (Bouzar-Benlabiod, 2023). While the CNN model is used to extrapolate the important features and isolate the tumor through hierarchical abstraction and multiple layers, the CBR (Case Based Recognition) is unique here and performs the diagnosis.

The model first goes through a data cleaning stage, where ResNext is used to retain the most coherent features of a mammogram (Bouzar-Benlabiod, 2023). Then, by using a pre-trained CNN model (SE-ResNet), features deemed important were excavated and used for classification by the CBR system. Various CBR systems were tested, including Particle Swarm Optimization (PSO), a Neuro-Fuzzy Adaptive System, and an automatic CBR that uses logistic regression and decision trees. The proposed architecture outperformed other standard classification methods such as straight logistic regression and radial basis function (RBF) kernels with an accuracy of 0.938 and an F-1 score of 0.984.

Another method of using deep neural networks to classify digital breast tomosynthesis (DBT) mammograms has been developed and tested which uses transfer learning. Using a deep convolutional neural network (DCNN) already trained for a similar task, this method re-adapts the model to a new task by training it with new data in that field (Heang-Ping Chan, 2020). This is critical for models that have limited data in their specific field.

For this specific model, as DBT imaging is scarce, transfer learning was of interest. Two fully connected layers were added to the pre-trained AlexNet NN, and the data was trained in either a single stage (either the DBT or Mammogram was used), or in two stages (Mammogram followed by a DBT). A random subset of the Mammogram data was used, and was able to boost the DBT AUC from 0.63 to 0.82 for the single stage training, while the AUC was boosted from 0.76 to 0.88 (Heang-Ping Chan, 2020).

Various neural network models have been trained and tested (e.g., CNN-, DNN-, RNN-, DBN=, and AE-based applications) (Maged Nasser, 2023), with some performing better than others. At a high level, the deep learning methods used two methods to classify the cancer: multi-class and binary classification. Binary classification reached a top accuracy of 98.7% (using a CNN feature selection architecture), while multi-class reached a peak of 95.6% using the same architecture with a sub type classification. Other models were also tested, with a binary FFNN that used negative and positive classes for detection reaching a 98.3% accuracy (Maged Nasser, 2023).

4 Data Processing

The data processing for our project involved several key steps, beginning with the sorting and conversion of image files. The DDSM (R. Sawyer-Lee, 2016) files were meticulously organized into folders based on their BI-RADS scores, with the assistance of accompanying spreadsheets. These files, originally in DICOM format, were converted into JPG format for ease of processing and uniformity. In contrast, the KAUMD (Asmaa S. Alsolami, 2020) files were already sorted by BI-RADS score, which streamlined the initial processing stage.

Following this, we combined the sorted image sets from DDSM and KAUMD into a unified dataset. A significant part of our preprocessing involved a custom python script, which cropped each image to focus only on the region of interest. This process was complemented by padding to ensure that the dimensions of each image were equal, thus creating uniform square images. We then resized all these images to a standard dimension of 224x224 pixels, ensuring consistency across the dataset.

One of the challenges we encountered was the skewness in the dataset's composition. To address this, we augmented specific subsets of the dataset, which was instrumental in enhancing the accuracy of our models and reducing bias. The final dataset was partitioned into training, validation, and testing subsets, with margins of 70%, 20%, and 10% respectively. This split was crucial for a robust training and evaluation of our models.

For demonstration purposes, we also included the INbreast (Inês C Moreira, 2012) database in our project. The INbreast files, like the DDSM files, were in DICOM format and underwent the same processing procedure, which included conversion to JPG format and the cropping and padding procedure. A particular focus in this phase of the project was the challenge posed by the varying resolutions of images within and between classes, especially in the KAUMD dataset. To tackle this, we developed a tailored approach to adjust image dimensions based on their original resolutions. This method aimed to minimize information loss during preprocessing. We identified the outermost pixels in every direction, recalculated ratios and resolutions, and adjusted dimensions accordingly.

5 Architecture

To excellently complete our BIRAD classification task, our primary model's architecture is using a stacked architecture which incorporates two pre-trained models, ResNet50 and GoogLeNet due to their depth and ability to learn complex features from images. In both pre-trained models, the final fully connected layer (originally designed for 1000-class ImageNet classification) is replaced with a new layer that has 5 output units, aligning with the BIRAD categories. The baseline model is used as a way to average the output of the stacked model.

During the training process, batch normalization is consistently applied to accommodate the monochromatic nature of the input images. Given the uniform color scheme, it becomes crucial to determine the optimal normalization parameters. To achieve this, the mean and standard deviation are computed across the entire dataset. The calculated mean for this dataset is 0.1152, and the standard deviation is 0.1977. These values are instrumental in standardizing the data, ensuring that the model processes the images effectively and consistently.

During the forward pass, an input image is passed through all three models (ResNet50, GoogLeNet, and the BaselineModel). Then the outputs of these three models are then averaged. This means each model contributes equally to the final prediction. Averaging helps in improving the robustness of the model, as it combines the strengths of each individual model and also prevents overfitting on the large CNN models.

Leveraging pre-trained models aids in robust hierarchical feature extraction, with higher-level modules being task-specific, thus efficiently recognizing low-level features. FIGURE 6 illustrates the layer configuration and data flow of Resnet. Thanks to the effective incorporation of branched convolutions within the residual modules, the model achieves efficient processing of multi-level features. (He, 2015).

Using GoogleNet allows the model to capture information at various scales and complexities due to the inception blocks and different size of kernels used in the model.FIGURE 7 illustrates the layer configuration of GoogleNet. The ability to capture features at various scales is particularly beneficial for mammograms, where both fine details (like microcalcifications) and larger patterns (like mass shapes) are important for accurate classification.

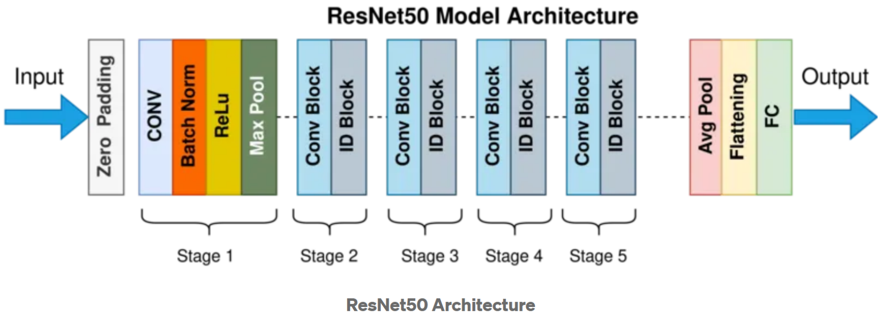


Figure 3: ResNet Achitecture

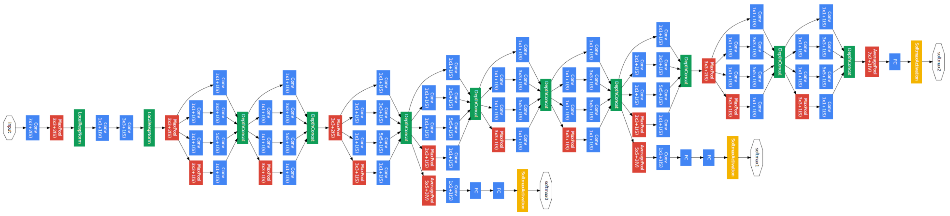


Figure 4: GoogleNet Achitecture

6 Baseline Model

The team decided to use the basic convolution neural network, depicted in Figure 5 as the baseline model.

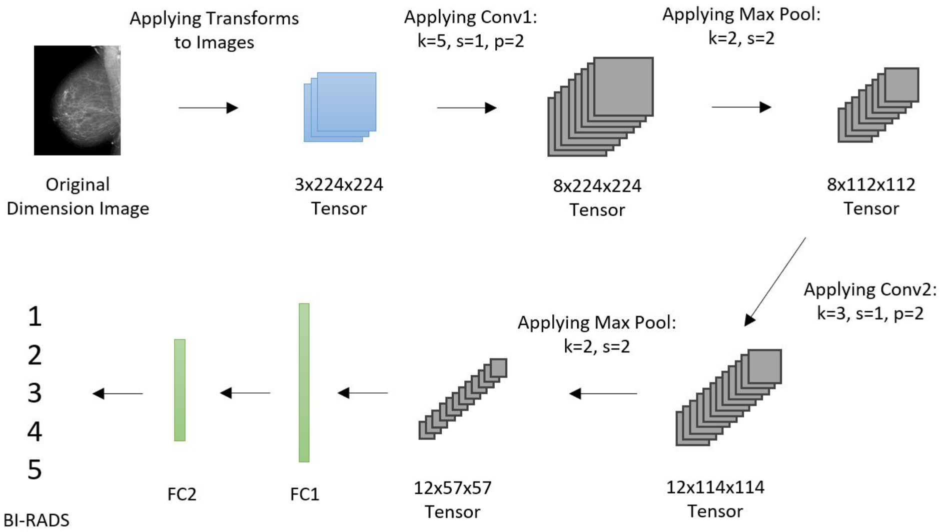

Applying Transforms to Images
Applying Conv1: k=5, s=1, p=2
Applying Max Pool: k=2, s=2
Original Dimension Image
3x224x224 Tensor
8x224x224 Tensor
8x112x112 Tensor
Applying Conv2: k=3, s=1, p=2
Applying Max Pool: k=2, s=2
1
2
3
4
5
BI-RADS
FC2
FC1
12x57x57 Tensor
12x114x114 Tensor

Figure 5: Conceptual diagram of baseline model architecture. Mammogram image: King Abdulaziz University mammogram dataset

The baseline consists of two convolutional layers, one with kernel size of five and another with kernel size of three; both convolutional layers have padding of two and stride of one. After each convolutional layer, max pooling is applied to reduce spatial dimensions by half. Following the second max pooling layer, the resulting feature maps are flattened to a vector which is then passed through two fully connected layers. These fully connected layers effectively map the features extracted from the convolutional layers to the five output classes. The chosen activation function applied after each layer (except for the second fully connected layer) is ReLU, as it is a standard activation function. The model is to be trained using cross entropy loss [nn.CrossEntropyLoss()] and the Adam optimizer.

7 Quantitative Results

7.1 Baseline Quantitative Results

The baseline model underwent several iterations of hyperparameter tuning, where the main hyperparameters investigated were batch size, learning rate, and number of epochs. The best performance on the validation set (without signs of overfitting) occurred with the following hyperparameters: batch size = 256, learning rate = 0.0001, number of epochs = 20; this set of hyperparameters was subsequently used during testing of the baseline model, producing the training and validation curves in Figure 6.

The validation accuracy of the chosen baseline model was approximately 63%, while the testing accuracy achieved was also approximately 63%. A confusion matrix was also developed to visualise the baseline's performance on particular classes during testing; as seen in Figure 7, the model performs best on BI-RAD classes one and four, followed by BI-RAD class two.

The overall precision, recall, and f1 scores for the baseline were determined to be 0.5833, 0.5943, and 0.5671, respectively.

7.2 Primary Quantitative Results

After training and hyperparameters tuning, the best model reached a training accuracy of 92% and validation accuracy is 72% . The detailed curve is shown in Figure 8. We see a fast convergence at the beginning of training due to the relatively big learning rate and a small batch size of 64. The model is performing well in the test set with an accuracy of 65.8% . To get a more intuitive insight

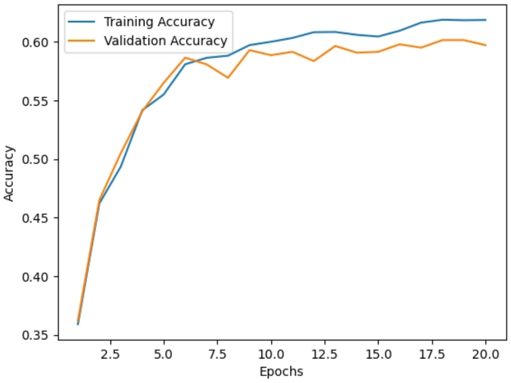


Figure 6: Training and validation curves for baseline model

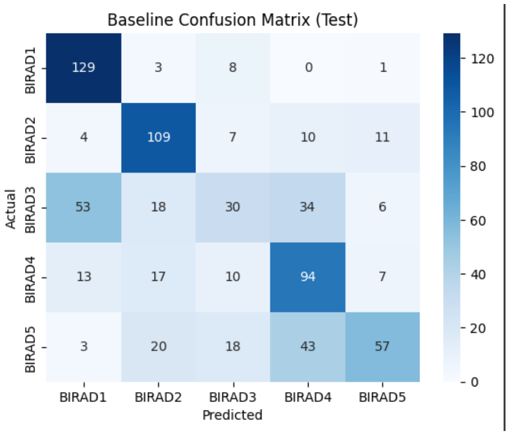


Figure 7: Confusion matrix for baseline model

on how the model performed, a confusion matrix is generated to help us understand what the model is predicting and contrast the performance on different stages of breast cancer. As shown in Figure 9, in general the model perform as we expected, however the model is having trouble correctly identifying BIRAD4 and BIRAD5. It will be discussed in more detail in the next section.

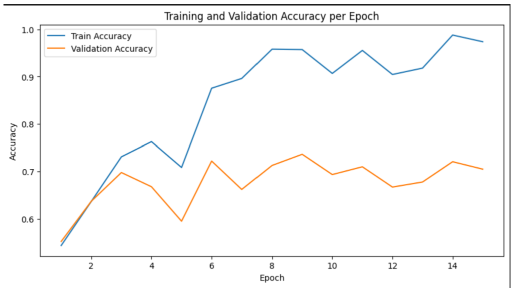
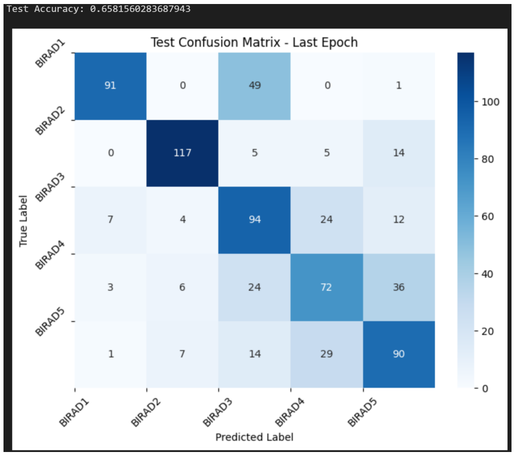


Figure 8: Training and validation curves for primary model



```
Test Accuracy: 0.6581560283687943
```

Figure 9: Confusion matrix on the testing set

8 Qualitative Results

8.1 Baseline Qualitative Results

With reference to the baseline confusion matrix (Figure 7), it is evident that the model is best at classifying BIRAD classes 1, 2 and 4. The performance on BIRAD classes 3 and 5 were significantly poorer. In particular, images for BIRAD Class 3 were most often classified as being BIRAD 1. This means that the model is classifying mammograms with signs of abnormality as non-malignant/normal scans, which can be very dangerous in a medical setting where the model is being used to aid a physician's decision making. This may indicate that the baseline is extracting similar high level features for classes 1 and 3.

8.2 Primary Qualitative Results

The confusion matrix provides insight into the model's performance, particularly highlighting the challenges it faces in distinguishing between BIRAD4 and BIRAD5 classes. We've observed that the model occasionally misclassifies images that, while having similar outer shapes, differ in internal patterns. This issue is compounded by the inherent difficulty in differentiating between BIRAD4 and BIRAD5 images, a task that is challenging even for trained human eyes due to their close resemblance.

Moreover, the overfitting of the training data is another concern. This appears to be partly due to the presence of watermarks on the images, which are there because the data is intended for research purposes and is publicly available. As shown in Figure 10, these watermarks inadvertently become features that the model learns to associate with specific classes, particularly if certain watermarks are more frequent in images of one class over another. Thus, the model might be making predictions based on the watermark patterns rather than the medical imagery, which is not ideal for accurate classification.

9 Evaluation on New Data

To further assess and evaluate our model's performance, utility and reliability, we tested our model on a brand new dataset, the handcrafted dataset. This dataset contains the same 5 classes as before to ensure the consistency of the classification. However, each class contains 100 images utterly different from any previous data seen by our models. The construction of this new set is described in the last paragraph of the Data Processing section. The new test set has no watermarks and has the cleanest data the team was able to find which is very close to a real-world modern mammography and it can test the model's ability to classify real-life medical-imaging on BIRAD levels. As for the results, our primary model achieved an accuracy of as high as 58% while facing such mammography images with no noise disturbance. This is similar to our validation but lower since there are no noise and exceeds our expectation because the graph looks pretty different with our train data as shown in conceptual diagram of Figure 10 and Figure 11. This demonstrates the complexity and difficulty of classification medical imaging in real life, and the reliability of our primary model.

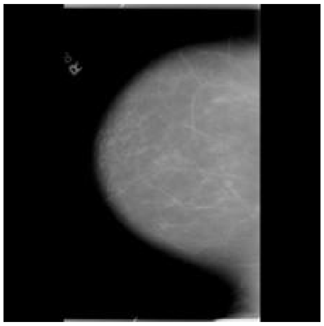


Figure 10: Picture from old set BIRAD5

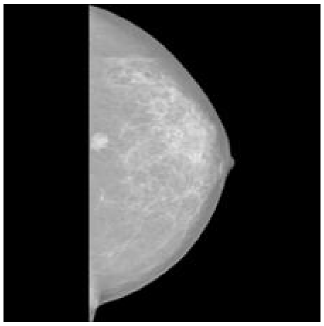


Figure 11: Picture from new set BIRAD5

10 Discussion

In recent years, stemming from revolutionary discoveries, both residual and inception learning models have transformed the landscape of image classification since their inception. By stacking both into the primary model, through ResNet 50 and GoogLeNet respectively, along with the team's pre-existing baseline architecture, it was hypothesized that a large boost in performance would present itself. Residual and inception learning boost performance using different methods:

1. ResNet-50 (residual learning) has played a pivotal role in improving model efficiencies, through the expulsion of vanishing gradients that stunted - or in some cases completely halted learning. (Sharma, 2023)
2. GoogLeNet (inception learning) has pushed boundaries of image classification efficiency through use of 1x1 convolutional filters to reduce dimensionality before applying larger convolutions. (Gomede, 2023)

Looking at Table 1, the model's accuracy improved by 7% (65% in the primary model compared to 58% in the baseline), while the overall f1 score increased from 0.5671 to 0.6 - -. Through integration of both GoogLeNet and ResNet-50, the increase in performance was tame, and less than the team expected. The integration of these two complex architectures raises questions regarding if the models are fully compatible and being used to their full capacity. The contrasting design principles may be contrasting one another, prompting the model to produce a non-trivial medium that sacrifices optimal tuning of each model, limiting final performance. Each was used independently to no avail, with the models not being able to overfit the data.

Table 1: Model comparison

Model	**Accuracy**	**F1 Score**
Baseline	58%	0.567
Primary	65%	0.6

Developing the optimal parameters for a mix of both models was challenging, with the team trying to balance computational cost, training accuracy, validation accuracy, and training loss. The team found the learning rate to have the most substantial effect on the results, as a low rate prompted extended training time, while a high rate overfitted the data and produced low validation accuracy. Some configurations produced high false-positive levels, some showed high false-negative values, and a couple configurations resulted in a reduced accuracy. In the context of medical imaging, producing a false-positive is almost always the more preferable of the two. A false positive will lead to unnecessary testing and scare, while false negatives can be fatal. To ensure the model is safer in a medical setting, hyperparameters were chosen that limit false negative classification.

While tuning, however, trends began to form that affected the model's ability to reach its highest performance. Primarily, this was both the inability to distinguish between BIRADS 1 and 3 and distinguish between BIRADS 4 and 5. Through tuning, these errors were minimized, with an emphasis on removing false-negatives, which was done for the confusion between BIRADS 1 and 3. The errors remaining in the model after tuning and training can be considered non-threatening, given the scarcity of false negatives.

An interesting insight into the model's performance can be tracked to a small feature present in one of the two training data sets. Following data clearing and preliminary training, the team noticed small watermarks in the data, appearing for BIRADS 2-5, but not 1. This led to predisposition for the model, as it now had a feature to grasp on to for specific classes - a feature that the model believed was part of the mammogram. This was an oversight by the team, as the model's performance would likely benefit from unmarked data. As the data had already been processed, and as a result of limited access to large sets of labeled mammograms, this hypothesis was not tested by the team.

The team cannot confidently recommend the model be used in practice. This starts with the fact that a trained professional would outperform the model and ends with the fear that the model can create confusion and doubt among medical experts if their classification does not align with that of the model. Through further architecture tuning and training, the team believes that this model can be further improved to a point where it may used as a tool to assist a trained professional.

As this was the first time working with deep learning for all team members, the posed task was a challenge. The team continually developed and scrapped models in the process of tuning and optimizing the model presented. From the baseline's simple CNN architecture, to residual learning and GoogLeNet used in the primary model, the majority of concepts applied stemmed from new knowledge developed in APS360.

11 Ethical Considerations

From an ethical consideration standpoint, the data collection process should be thoroughly reviewed so as to ensure diversity in patient body types, for example. It has been noted that women with a high body mass index (BMI) receive screening mammograms less often in comparison with women with lower BMIs (Wee et al., 2004). Due to having fewer screenings, it is possible that existing data sets containing mammograms are less representative of women with a high BMI. A model trained on such data may cause a bias toward women with lower BMIs, and perform poorly on scans of someone with a high BMI.

It must be disclosed that the model cannot be relied upon as the only source of direction when it comes to deciding the patient's care; as well as the model may perform during testing, it must only be used as an aid to the physician, as it can be difficult to ensure that the data set used to train the model would be able to account for pathology outside of what is typically seen in individuals with breast cancer (i.e., patients who have rare forms of breast cancer that are likely not present in the training data). Fundamentally, given the consequences of a misdiagnosis by the model, it is essential that a physician thoroughly review each patient's scans and make a final decision on the patient's condition.

12 Project Difficulty & Quality

Developing a deep learning model for the analysis and categorization of mammograms is a challenging and complex task that is, to this day, continually being researched and improved. While improving, many current models have shown performance inconsistencies on unseen data that limits their in-lab use. Models are trained on hand-crafted datasets to maximize accuracy, but begin to fall short of expert classification when shown external data. This raises concerns regarding the strength of model generalization and validation levels for most models. (Wang, 2020)

The team experienced this in both the training and testing of the model, as multiple data sets were used in both stages. To train the model on all 5 BIRAD levels, data from each was required, but a dataset satisfying this could not be found. Discussed in section 4, this was resolved by combining two datasets, which indirectly introduced inconsistencies in the imaging between BIRAD levels. To fully test the model it was also required that an external dataset was used, featuring data from an unknown, external set. In the training, it can also be noted that BIRADS 2-5 from the first dataset had watermarks that may have had a slight impact on the tuning.

The model also developed classification trends that hindered the models performance, namely confusing BIRADs 1, 3 and BIRADs 4, 5:

BIRADS 1 and 3

1. Between BIRADS 1 and 3, the main source of error comes from a misclassification of BIRAD 1, classifying it as BIRAD 3 (49 BIRAD 1s were classified as 3, while only 7 BIRAD 3s were classified as 1). Looking back to the data processing, one of the two datasets used consisted primarily of BIRADs 1 and 3,

BIRADS 4 and 5

1. For BIRADS 4 and 5, there are classification issues both ways. This is likely a result of the similarities between the levels, as miniscule features differentiate them, a task that troubles even the trained human eye.

Given the project difficulty, the model performed adequately, underperforming professionally developed models, but outperforming the baseline. (Wang, 2020) Issues that commonly arise in mammogram classification models were still present in the team's primary model; specifically the poor performance on data from a new source.

As a result of the project's difficulty, the primary model was continually changing throughout the entire project. Numerous architectures were studied, with the model starting as a simple residual learning (ResNet-50 model), which was changed to inception learning (GoogLeNet), and finally back to a combination of both. The final model was unable to train on the team's laptops, and a special GPU was required, which was provided by a team member. While the performance was not up to the standard set by industrial models, significant attempts were undertaken by the team to maximize performance. While not perfect, the model produced an accuracy over 80% and f1 score greater than 0.5 for each BIRAD, signifying that the model is performing the task given to it.

13 GitHub Repository

Our work will be done on GitHub. The following link leads to the repository: https://github.com/yazeedbukhari/BIRAD-classification

REFERENCES

Sarah Eskreis-Winkler Arka Bhowmik. Deep learning in breast imaging. *National Library of Medicine*, 2022. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9459862/.

Wafaa Alsaggaf Sawsan Ashoor Haneen Refaat Mohammed Elmogy Asmaa S. Alsolami, Wafaa Shalash. King abdulaziz university mammogram dataset, 2020. URL https://www.kaggle.com/datasets/asmaasaad/king-abdulaziz-university-mammogram-dataset.

Lydia Bouzar-Benlabiod. A novel breast cancer detection architecture based on a cnn-cbr system for mammogram classification. *Science Direct*, 2023. URL https://www.sciencedirect.com/science/article/pii/S001048252300598X.

Canadian Cancer Society. Screening for breast cancer, 2023a. URL https://cancer.ca/en/cancer-information/cancer-types/breast/screening.

Canadian Cancer Society. Breast cancer statistics, 2023b. URL https://cancer.ca/en/cancer-information/cancer-types/breast/statistics.

MD Daniel Liu. What are birads categories?, 2022. URL https://www.cancercenter.com/cancer-types/breast-cancer/diagnosis-and-detection/mammography/results-bi-rads#:~:text=The%20test%20results%20are%20scored,four%20categories%20for%20breast%20density.

Macedo Firmino, Giovani Angelo, Higor Morais, Marcel R. Dantas, and Ricardo Valentim. Computer-aided detection (cade) and diagnosis (cadx) system for lung cancer with likelihood of malignancy. *BioMedical Engineering OnLine*, 15(1), 2016. URL https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/s12938-015-0120-7#:~:text=CADe%20are%20systems%20geared%20for,between%20benign%20and%20malignant%20tumors.

Everton Gomede. Exploring googlenet: A revolutionary deep learning architecture. 2023. URL https://medium.com/@evertongomede/exploring-googlenet-a-revolutionary-deep-learning-architecture-8bb176a0facc#:~:text=GoogLeNet's%20impact%20on%20the%20field,advancements%20in%20neural%20network%20design.

Ramzi Guetari, Helmi Ayari, and Houneida Sakly. Computer-aided diagnosis systems: A comparative study of classical machine learning versus deep learning-based approaches. *Knowledge and Information Systems*, 65(10):3881–3921, 2023. URL https://link.springer.com/article/10.1007/s10115-023-01894-7#Abs1.

Kaiming He. Deep residual learning for image recognition. 2015. URL https://arxiv.org/abs/1512.03385.

Ravi K. Samala Heang-Ping Chan, Lubomir M. Hadjiiski. Computer-aided diagnosis in the era of deep learning. *National Library of Medicine*, 2020. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7293164/.

Inês Domingues António Cardoso Maria João Cardoso Jaime S Cardoso Inês C Moreira, Igor Amaral. Inbreast: toward a full-field digital mammographic database, 2012. URL https://pubmed.ncbi.nlm.nih.gov/22078258/.

Umi Kalsom Yusof Maged Nasser. Deep learning based methods for breast cancer diagnosis: A systematic review and future direction. *National Library of Medicine*, 2023. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9818155/.

A. Hoogi D. Rubin R. Sawyer-Lee, F. Gimenez. Curated breast imaging subset of digital database for screening mammography (cbis-ddsm) [data set], 2016. URL https://doi.org/10.7937/K9/TCIA.2016.7O02S9CY.

Saleem Z. Ramadan. Methods used in computer-aided diagnosis for breast cancer detection using mammograms: A review. *National Library of Medicine*, 2020. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7091549/.

Tanish Sharma. Detailed explanation of resnet cnn model. 2023. URL https://medium.com/@sharma.tanish096/detailed-explanation-of-residual-network-resnet50-cnn-model-106e0ab9fa9e.

U.S. Food and Drug Administration. Mqsa national statistics, 2023. URL https://www.fda.gov/radiation-emitting-products/mqsa-insights/mqsa-national-statistics.

Xiaoqin Wang. Inconsistent performance of deep learning models on mammogram classification. 2020. URL https://www.sciencedirect.com/science/article/pii/S1546144020300284.

Christina C. Wee, Ellen P. McCarthy, Roger B. Davis, and Russell S. Phillips. Obesity and breast cancer screening. *Journal of General Internal Medicine*, 19(4):324–331, Apr 2004. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1492197/.