Deep Learning Based Data Augmentation for Breast Lesion Detection

Zhendong Yuan

CMU-CS-21-129

Aug 2021

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Adam Perer, Chair
Zachary Lipton

Submitted in partial fulfillment of the requirements for the degree of Master of Science.

Keywords: Breast lesion detection systems, data imbalance, SMOTE, up-sampling, down-sampling, image processing, ResNet, GAN, deep learning

Abstract

Deep learning has become increasingly popular in a wide range of applications in the past few years. The performance improvements in hardware and machine learning models have made it possible to train a deeper and wider network to achieve state-of-the-art (SOTA) performance for those applications. However, there still exist several potential obstacles that researchers have to overcome before producing a model that could actually be useful in reality. One of the common obstacles is related to the data itself. The training data collected from a small hospital could be limited in quantity and a pre-trained model taken from other hospitals could have bad generalization performance due to potential differences in the X-ray machines and the environment in which the mammogram is taken[41]. Moreover, since the majority of the data collected from the mammogram comes from patients who actually have no illness, there could be a serious imbalance of positive/negative cases in the training data. Models trained using such data could naively achieve an extremely high overall accuracy by predicting everything as normal and would have no actual value in reality. However, lesion/cancer detection is a task that requires the model's predictions to be accurate for both positive/negative cases, resilient to noises, and consistent across different data sources.

In this thesis, we provide workarounds for the issues mentioned. Our experiment is based on the UPITT mammogram dataset that is comprised of 79501 images collected from approximately 22267 distinct patients. In order to deal with the dataset size restriction and to achieve localized explanation, we decide to use a patch-based model for the lesion classification. We extract the normal patches from the breast tissue in images with BIRADS level of 1. The lesion patches are extracted from the ROI(region of interest) labeled by the radiologist from images with BIRADS level score of 0,2 using computer vision techniques. We designed our own techniques to deal with the serious data imbalance via deep learning-based SMOTE[9] and GAN[6, 12, 18, 28] and test those techniques with a deep convolutional model that is similar to VGG16[35].

Acknowledgments

I would like to thank my advisor, Adam Perer, for providing me with the access to the UPITT-mammogram dataset and guidance on the research direction and helping me resolving the concerns and obstacles that I encountered. I would like to thank professor Zachary Lipton, for offering me assistance over machine-learning related questions and providing the GPU machine for development and model training. I would like to thank Sachin Grover for his guidance on image processing and Bryan Lu for his assistance on dataset statistics collection.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Breast cancer is a type of cancer that forms in the cells of the breasts and it is one of the most common cancers diagnosed among women in the United States[1]. According to the statistics from the BreastCancer.org[3], about 1 in 8 US women will develop breast cancer over the course of her life. In 2021, there are an estimated 281,550 new cases of invasive breast cancer expected to be diagnosed in women in the US along with 49290 new non-invasive cases. On the other hand, the statistics for men are 2,650 new invasive cases and 1 in 833 men will develop breast cancer in his lifetime.[3] Because of its high death rate among women, the disease has raised substantial awareness and gathered much research funding and support to help with its diagnosis and treatment.

Screening Mammograms, an x-ray of the breast, are routinely administered to detect breast cancer among patients who have no apparent symptoms of breast cancer.[2] If the radiologist finds the result of screening mammograms suspicious, what usually follows is either a diagnostic mammogram or ultrasound. Therefore, the reliability of screening mammogram is critical as it is usually one of the first stages of breast cancer diagnosis. Its reliability depends on several factors such as the size of the tumour, the density of the breast tissue and the skill of radiologist who reads the mammogram. Under most of the situations, the radiologist looks for certain suspicious patterns in the mammogram such as calcification, masses and soft tissue lesion. According to statistics, after timing seven radiologists as they interpreted a total of 181 digital and 183 film-screen screening mammograms, the average amount of time that they spent on the digital studies is 2.0 minutes.[7] Thus, there could be significant value in provision of certain form of assistance when a radiologist is making diagnosis under such a restricted time-frame as the decision could potentially be a matter of life or death.

Deep learning is probably one of the biggest achievements/breakthroughs in the field of machine learning in the past few years. It has spawned a wide range of applications in different fields such as natural language processing, object detection, facial recognition, image classification and medical diagnosis. As the hardware performance keeps improving, deeper, wider neural networks now have the potential to achieve human-level of accuracy even in difficult tasks like translation and medical diagnosis as training cost significantly lowers. This means that deep learning models could potentially be a great assistance for radiologists to improve their accuracy

and save the time in the near future.

In order to produce a practical and applicable model to assist the radiologists, we believe that there are three major obstacles that need to be overcome. Specifically, they are a proper way of handling the fundamental restrictions imposed by the data itself, development of a model architecture that is suitable for the size of the data and properly trained using the right set of hyperparameters such that it could achieve reasonable generalization performance, and a localized prediction for a given area of the given mammogram instead of an overall prediction for the mammogram so that the model could actually serve as an assistance to the radiologist in reality.

Similar to most image classification tasks, medical diagnosis is heavily dependent on the quality and quantity of the input images. One of the most common approaches is to use a CNN-based architecture trained using the labeled mammogram x-ray images directly. This approach has its own merit if the quantity of data is sufficiently large and the quality of the images is consistently high. However, in certain fields such as breast cancer detection, there are several important concerns that differentiate it from a task like melanoma detection. Due to privacy concerns, the quantity of data available could be significantly lower for breast lesion detection. Moreover, as majority of the patients who underwent mammogram scanning are negative cases, there is a serious imbalance in the quantity of positive/negative data which could cause serious model performance degradation. This is the reason why data augmentation is so critical in achieving better performance than a model trained on the original dataset.

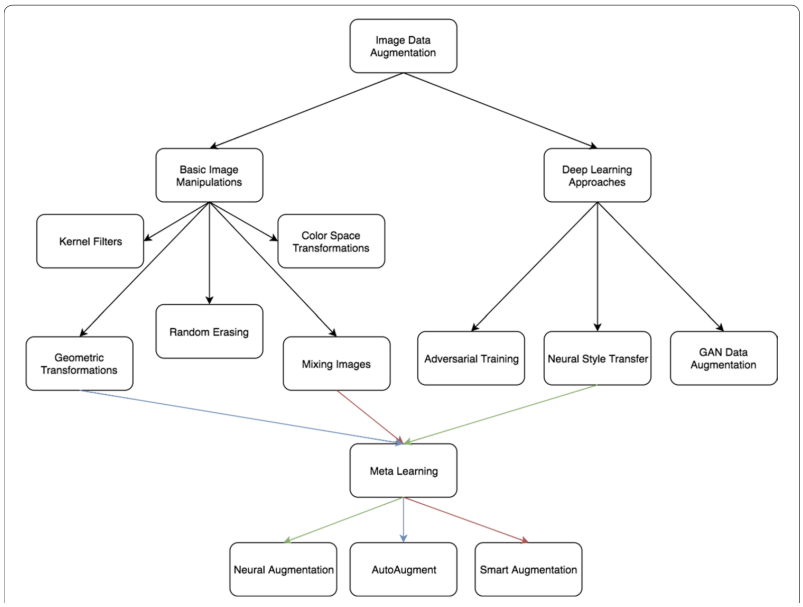


Figure 1.1: Data Augmentation Techniques Overview from [34]

The two standard routes for data augmentation are the image-manipulation based approach and deep learning based approach. The deep learning based approach became popular in the last few years and took advantage of deep convolutional layers for extraction of the features in the

images. One of those architectures called GAN has the capability of taking a vector in the latent space and generate data that is unseen in the original dataset. This approach allows the model to observe some data points that are not existing in the training dataset and potentially achieve better test performance. Since then, lots of research interest was gathered on finding a proper evaluation metric to optimize GAN and determine the extent to which GAN could help improve model performance if the GAN is actually trained on the same dataset as the classifier. [5, 6, 12, 28]. In a liver-lesion detection study, GAN-based data augmentation improved the model performance by increasing the sensitivity/recall and specificity/precision from 78.6% 88.4% to 85.7% and 92.4% respectively[34]. Therefore, it could be possible that the GAN would also help improve classifier performance even when it is trained on a seriously imbalanced dataset. On the other hand, image-manipulation based approach's major advantage is that it is simple and straightforward as it applies some operations on the pre-existing images directly. Simplest methods are using geometric transformations such as vertical and horizontal flipping of the original image to slightly expand the dataset size and introduce some variations. Other up-sampling technique, such as SMOTE that was created 20 years ago to tackle data-related problems, basically creates a mixture of input data based on interpolation from nearest neighbors of a base vector using some pre-defined distance metric. [9]

Furthermore, there could also be significant qualitative differences in the mammogram if it is taken with a different machine under different views of the breast or due to the differences in the size of the breast or tissue density. Also, as the ROI (region of interest) in a mammogram is expected to cover a significantly smaller proportion of the whole mammogram compared to the x-ray of other disease such as melanoma that is usually taken for a small target skin area, it is critical that we adopt certain image processing or data augmentation technique to eliminate the potential sources of noises to make sure that the model's performance is more consistent across different data sources that is different from the training data-set. One of the approaches would be, instead of training the a deep-learning model on the whole mammogram, we divide the image into patches and try to predict whether each patch is actually representative of some forms of lesion so that the variations across training data is reduced. This approach is based on the assumption that the judgement made by the radiologist is usually based on a relatively small localized area rather than a holistic view of the whole breast since it is unlikely that two distant patches viewed together would provide the radiologist with additional information in determining whether one of them is a lesion.

However, having an accurate model still isn't sufficient for it to be applicable in the industry. When we rely on a model that is trained using the whole mammogram images, interpretation techniques such as LIME[29] need to be applied so that some regions could be highlighted by the activations during the explanation phase. However, in our use case, as we broke the mammogram image into much smaller patches, we could directly tell the pattern that the model is recognizing from the prediction result and provide localized explanation for the radiologist without extra efforts.

In recognition of the challenges in building a breast lesion detection system, we provide a complete solution that works for general mammogram datasets. The solution is comprised of an

algorithm for image processing to extract patches representing the ROI (region of interest) from the mammogram, deep learning based techniques that we designed to tackle the severe data imbalance in medical image analysis, development of a CNN-based architecture that makes the classification of unsuspicious/lesioned patches, and an attempt to provide an explanation for the model perception/behavior using localized information based on the patch-based model.

In this thesis, we make the following contributions:

- Implement a generic image processing, patch extraction logic for mammogram dataset.
- Implementation of a CNN-based model for the experimentation with the UPITT mammogram dataset and treat it as the baseline for model performance evaluation.
- Implement deep learning-based SMOTE that relies on interpolation in the hidden layer/feature layer and GAN for data augmentation.
- Provision of patch-based localized prediction using the models developed.

Chapter 2

Background and Related Work

We next provide background on the current state of the art approach to cancer/lesion classification in medical image analysis and a brief analysis of the common issues that cause bad generalization performance of the classification models trained on a specific dataset. Then we discuss some latest techniques that were applied in some other classification tasks to deal with the dataset issue and consider their trade-offs.

2.1 State of the art model

2.1.1 General architecture

Despite its prevalence in a lot of applications in the past few years, convolutional neural network has received little attention in the medical field until recently.[25, 33, 40, 41, 42] This is attributed to the fact that data is always one of the most crucial parts of machine learning and is quite indispensable for achieving a reasonable model performance. A common CNN architecture consists of multiple convolutional layers, pooling layers, and linear layers stacked on top of each other. Each convolutional layer mainly consists of a set learnable filters that extract the features from the input. The pooling layer basically reduces the dimensionality of its input by combining the outputs from a cluster of neurons in the previous layers. The fully connected layers are the multi-layer perceptrons that classify the images based on the feature maps constructed from previous feature extraction layers. In the last few years, several different architectures, e.g. VGG, AlexNet, ResNet, have been proposed.[15, 21, 35] Despite of the intrinsic similarities among those architectures, their performance could actually have significant differences and there have been a large number of literature comparing the performances of them on different image classification tasks. For most of the cases, a deep and wide ResNet architecture with a large number of trainable parameters could have better generalization performance on a complex classification task. However, this also results in longer training period and cost, and sometimes it could be an overkill and the performance could even be worse than a much smaller network that does not overfit to the training dataset when the training samples do not have that high resolution.

2.1.2 Mammogram-image based classification vs Patch-based classification

Two common approaches for breast lesion/cancer classification are mammogram-image-based and patch-based and next we will examine their advantages and disadvantages respectively. The mammogram-image based method usually involves simple pre-processing that extracts a fixed window that contains the breast tissue. The patch-based method involves more complex preprocessing logic for the image based on the BIRADS level for the image. The normal patches are extracted by using a similar approach as the extraction of the breast region in the mammogram-image-based classification with an additional step that iterates over the breast in blocks of fixed size and applies some criteria for filtering and selecting high-quality patches. Lesion patches are extracted from the ROI(region of interest) annotated by the radiologist and the process could involve several steps of image preprocessing before we eventually crop the lesion patch from the annotated region.

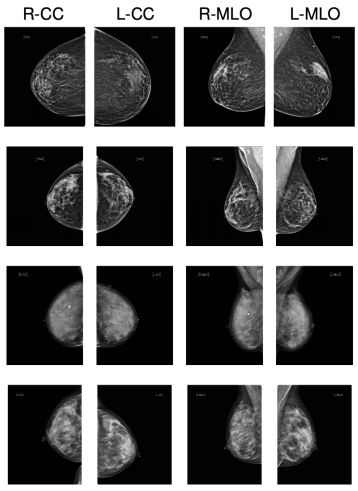


Figure 2.1: Standard Views Required for Screening Mammogram and View-based Classifier[25, 41, 42]

The mammogram-image based approach seems to be the more intuitive approach for lesion classification. The clear advantage of mammogram-image based classification is that the image processing would be much simpler and the labeling could also be extracted from the reports directly. However, if the quantity of the training data is limited, the mammogram-image based model trying to extract the features from the whole breast could have poorer performance due to more noises in a global view, e.g. differences in the angle and size of the breast, the density and distribution of the tissues. The model could be harder to train because it has to be able to recognize the lesion that only covers a really small area of the image. More importantly, the prediction from the model does not provide much information regarding how the decision is made, thus hurting interpretability.

In recognition of the first data-related issue mentioned above, a finer-grained view-based classification model have been proposed and it could be built using on an ensemble of models trained

for each view separately to achieve better accuracy than a model that does not distinguish across different views. As we could tell from Figure 2.1, by training the model on a fixed view of the breast, it effectively reduces the variation across the images and could improve the model accuracy and convergence speed. [25, 41, 42]

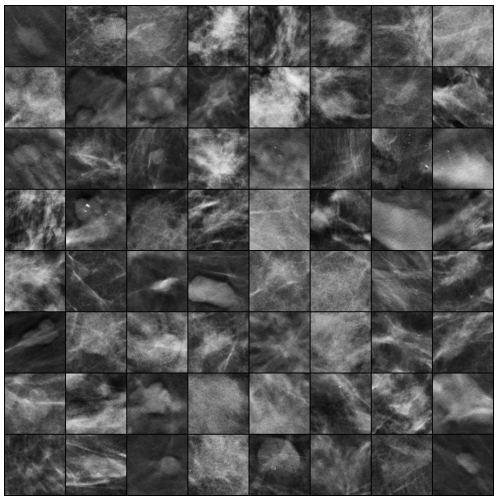


Figure 2.2: Sample patches extracted based on radiologist annotations

The patch-based approach requires that the suspicious region in the mammogram be annotated by the radiologist for patch extraction and label generation. It adds extra complexity and source of error in the data collection step. There are several major challenges in the data collection procedure.

- The decision on the patch_size(resolution) could significantly impact on the model performance. If the patch is too small, a single patch might not contain sufficient information that indicates the radiologists' findings. If the patch is too big, the model could have difficulty focusing on the actual pattern that is suspicious and have difficulty generalizing.
- The filtering procedure has to effectively eliminate the mammogram background and target regions of breast tissue.
- In the model training phase, the lesion patch must not contain any hints from the radiologist, meaning that the annotations from the radiologists must not be included to cause potential data leaking and model performance degradation.

Despite of those challenges that require tuning and application of some programmatic techniques, the patch-based method also has several clear advantages over mammogram-image based model.

- Significantly expands the dataset size. Effectively deal with the issue that the model underfits with too few mammogram images.
- Less variability across training datasets by focusing on a much smaller region of tissue instead of a global view that is sensitive to the dimension, view, rotation of the breast as long as the same patch extraction logic is applied and the radiologist annotation style is fixed. So the model could potentially have better generalization performance across different datasets.

- Ability to directly provide localized explanation that could be taken advantage of to assist the radiologist.
- The potential for a pre-trained patch-based classifier to be converted into a mammogram-image based classifier for images without ROI annotated. This is actually quite intuitive as we could basically reduce the problem of lesion prediction from mammogram image to prediction from patches by iterating through the mammogram image and making a final prediction based on localized predictions made. [32]

2.2 Data imbalance and data augmentation

One of the most prevalent issues in the field of medical image analysis is that the distribution of the labels is usually quite imbalanced, with majority of the training data being negative cases.[25, 41, 42] This leads to the problem that a naive model that actually has degraded performance could potentially get really high training/test accuracy by predicting everything as negative. Therefore, we will explore some of the approaches that are frequently applied in other fields such as forgery detection to discuss their applicability in breast lesion detection or other medical image analysis tasks. We will also take this into account when we are eventually evaluating our classifiers to ensure that our model carries actual significance in reality.

2.2.1 Naive up/down sampling

We first analyze the applicability of simple up/down sampling approach.

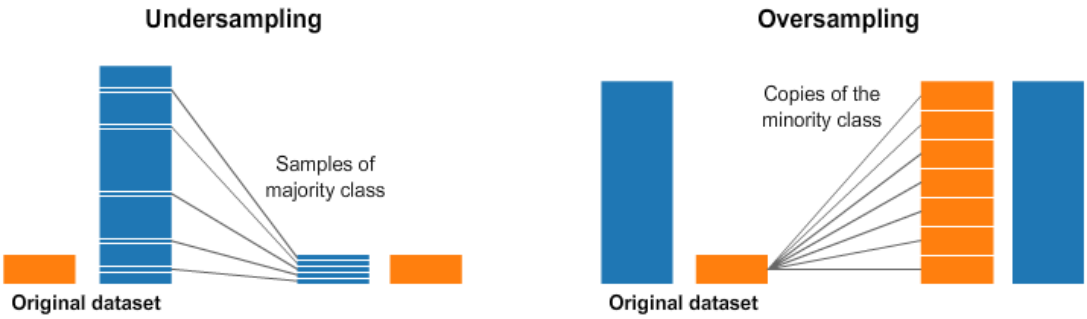


Figure 2.3: Data Over Sampling and Under Sampling Illustration from [20]

Under sampling

Under sampling is a non-heuristic method that aims to balance the distribution of the dataset by randomly eliminating samples from the majority class of the dataset.[20] This approach makes it possible to achieve arbitrary distribution of the samples within the dataset without the need for generation of new samples that are potentially inconsistent with real data. However, this mechanism also discards some data that could potentially be useful for the induction process. Under sampling could hurt model's ability to estimate the true distribution of target by interfering with the data sampling process. Furthermore, Under sampling might do little to help the model

achieve better generalization performance as minority class data is still limited so the model would not necessarily produce a more accurate decision boundary even if the sampling process enforces a 1:1 ratio between positive and negative cases. Essentially, the model still learns the same set of features for the minority/positive class after it is trained for a sufficiently large number of iterations. On the other hand, as some of the majority/negative samples are discarded in the sampling process, model could actually perform worse to create more false positive without decreasing the number of false negatives.

Over Sampling

Over sampling, is a non-heuristic method that aims to balance the distribution of the dataset by randomly replicating samples (with replacement) from the minority class of the dataset.[20] The approach could be effective in achieving balance of the dataset but the potential issue is that the approach still does not really present new data that does not exist in the original dataset to the model during the training process. After some epochs, the model would have seen all the samples from the minority class and the sampling technique's impact on the final model performance could be limited if no new minority class data is added. Given sufficient iterations of training, the model would still overfit to the negative/majority cases and experience performance degradation because it is not actually exposed to unseen features of positive samples in the sampling process.

2.2.2 Interpolation based over-sampling: SMOTE & ADASYN

SMOTE was introduced in the 2000s to deal with data imbalance issue.[9] It can be regarded as an over sampling approach that replicates the samples from the abnormal class. The author for SMOTE claims that the traditional over sampling with replacement only creates a more specific decision boundary that essentially overfits with the minority samples without improving the generalization performance because replication of samples do not make the minority decision boundary spread into minority class region. Therefore, the method they proposed is to use the k nearest neighbors for each sample in the vector space and generate new samples by using an interpolation from the original sample to its nearest neighbor by a random proportion/distance. Therefore, the new data points created for the minority class could potentially expand the decision boundary or make it more precise to allow the model achieve better performance on the test dataset it is evaluated on.

ADASYN[14] basically follows the same logic as SMOTE. The difference between ADASYN and SMOTE is that ADASYN introduces several coefficients that affect the number of data points interpolated from one base data point to its neighboring data points rather than using a fixed parameter k in the k-nearest algorithm to generate k data samples for each data point in the original training dataset to expand it by k-fold. ADASYN algorithm basically takes the type/class of the neighbors for a data point into account when it is deciding the number of samples to be generated for it. The coefficient r_i is computed as a ratio of the number of neighboring samples that belong to minority class over k, the number of neighbors examined, and then it is normalized across all minority data points and it determines the number of synthetic samples to be generated for each

data point.

$$k = \text{number of neighbors to be examined}$$
$$\beta = \text{ratio that controls the balance between number of samples}$$
$$G = (n_{majority} - n_{minority}) * \beta$$
$$r_i = \frac{\text{number of minority samples among k nearest neighbors}}{k}$$
$$\hat{r_i} = \frac{r_i}{\sum_i^{n_{minority}} r_i}$$
$$G_i = G\hat{r_i}$$

Theoretically, the SMOTE approach would lead to a larger and more precise minority class region that could potentially include some minority samples that have similar attributes but are unseen in the original dataset to improve model performance. However, the shortcoming for this approach is also quite obvious as there is no guarantee on the degree/direction of the minority class region expansion since a random coefficient is applied during the interpolation phase. Therefore, it could introduce a large amount of noise that actually worsens the model performance. Furthermore, the method was originally proposed for structured data that are already vectorized. This means that it is not suitable for direct application with image data and a workaround is necessary. This is the reason why we would like to explore the effectiveness of a modified version of this approach in this medical image analysis experiment and propose another method to compare their performances.

2.2.3 Deep learning-based over-sampling: DFBS and GAN

As the processing power of the hardware keeps improving, deep learning based approach for synthetic data sample generation gained its popularity in the last few years. The deep learning approach has its clear advantages over traditional replication-based or interpolation based over-sampling in several aspects. The feature extraction process is further optimized to ensure a more precise region in the feature space for each class. Here we will discuss two deep learning-based approaches DFBS and GAN that achieve these optimizations.

DFBS

DFBS, or deep discriminative feature-based sampling was proposed to deal with some problems with SMOTE.[22] Similar to SMOTE, DFBS tries to expand the class region of minority samples in the feature space. The difference is that DFBS pays more attention to the expansion process by taking the majority class into account to ensure that the generated minority sample are more distant from the majority class region than the minority class region by some predefined distant metrics.

Algorithm 1: DFBS Algorithm

Data: D^m: minority data, D^M: Majority data, ρ: oversampling rate
Result: F^s: synthetic minority feature vectors, F^m: minority feature vectors, F^M: Majority feature vectors

$D = D^M \cup D^m$
n_m = the number of minority class samples
n_s = the number of synthetic feature vectors
d = feature size of transformed feature vectors
model$_d$ = deep discriminative network
training model$_d(D)$
F^M = model$_d(D^M)$
F^m = model$_d(D^m)$
calculating three center points c^M, c^m, c^{all}
while $n_s < n_m * \rho$ **do**
    for $j = 0$ **to** d **do**
        f^s_j = randomly sample a value from the collection of f^m_j
    end
    if dist$(f^s, c^M) >$ dist$(f^s, c^{all}) >$ dist(f^s, c^m) **then**
        add f^s into F^s
        $n_s = n_s + 1$
    end
end
return F^s, F^m, F^M

Figure 2.4: DFBS algorithm from original paper [22]

The major principle behind this approach is that the pre-trained auto-encoder will encode the input into a fixed dimension feature vector. Then a new sample will be generated by a combination of existing feature vectors. In addition, one of the major contributions from the author is that they introduced this extra step of verification that measures the distance of the new sample from minority class center, majority class center, overall center to determine whether the generated sample should be considered as valid. This ensures that the generated minority sample won't over-extend into the majority class region and alleviate the instability issue that exists in SMOTE because of its random interpolation. However, there are also some potential flaws in the verification step that we should be concerned with.

- The verification step could be superfluous. The occasion that interpolation between two minority data points are closer to the centroid for the majority class region should be rare to have significant impact on the final performance.
- Closer distance in low level feature space does not necessarily translate to closer distance in high-dimensional space. This is one potential problem with all nearest neighbor and SMOTE based augmentation approach because the network would apply non-linear transformations to the low-level feature and completely change the original distance relationship among data points.
- Euclidean distance applied in the original DFBS algorithm is not necessarily a good metric when it is used for high-dimensional feature vectors as difference in one dimension could significantly impact on the final distance. This is the reason why some researchers propose the application of l1-norm or even fractional norm for high-dimensional data. [4] We will

explore some other distance metrics in this experiment.

GAN

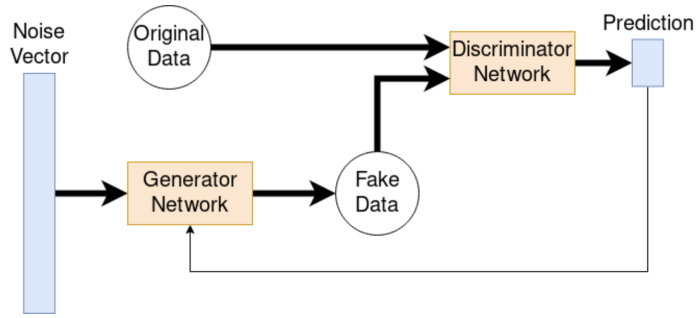


Figure 2.5: GAN Mechanism Illustration from [11]

GAN is another deep learning-based data augmentation approach that became increasingly popular in the last few years for its ability to learn the low-level representation of the data samples in the feature space and produce high quality fake data. [11] It was shown in previous research that GAN-based data augmentation could quite significantly improve the performance of image classifiers.[34] The major difference between GAN and DFBS is that GAN has a built-in discriminator that tries to learn the distinction between fake data and real data as opposed to the verification function in DFBS. The discriminator basically serve as a sanity check for data sample produced by the generator using a random noise vector. From Figure 2.6, the training process can be regarded as a zero-sum game in which the discriminator tries to tell the difference between fake data and real data and maximize the loss function and generator tries to fool the discriminator and minimize the loss function. This usually results in a much longer training process since two networks have conflicting interests. Also, this means that more regularization techniques have to be applied to prevent one network from getting too perfect such that the loss would basically vanish to stop the learning of the other network. In this paper, we will compare the performance of the classifier augmented with synthetic data generated from GAN with the baseline classifier that only applies image manipulation based augmentation.

$$\min_G \max_D V(D,G)$$
$$V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z))]$$

Figure 2.6: GAN minimax function

Chapter 3

Deep Learning Augmented Breast Lesion Detection System

We now describe our breast lesion detection system that is augmented with synthetic data from deep learning algorithms. Our goal is to implement an approach that effectively solves the common challenges faced by the traditional medical image analysis system. We will explore the potential of some deep learning algorthms to deal with the model's inherent challenges faced by majority of the medical imaging dataset.
In this experiment with the UPITT mammogram dataset, we will compare the performance of the baseline model with un-augmented data with models that are augmented with synthetic data that are generated with deep-learning based SMOTE and GAN.

3.1 Dataset

3.1.1 General Statistics

Dataset

Our mammogram dataset is provided by the University of Pittsburgh for academic and research purposes. The dataset consists of 101,494 DICOM images from 22,267 patients whose personal information was de-identified before we received the data. Each patient has a variable number of images and it is common that some of the views are missing.
The figure-3.1 summarizes the distribution of the number of images per patient across the whole dataset. On average, each patient has 4.56 images in the folder with the standard deviation being 4.14. The large standard deviation and the figure-3.1 indicates that most of the patients do not have their complete mammogram results documented, which means that view-based model in Figure 2.1 that requires all 4 views to be present becomes impractical in those situations. This is the reason why we choose to adopt a patch-based approach that is mostly unaffected by the completeness of data and less sensitive to the qualitative differences across the images compared with the mammogram-image based approaches.

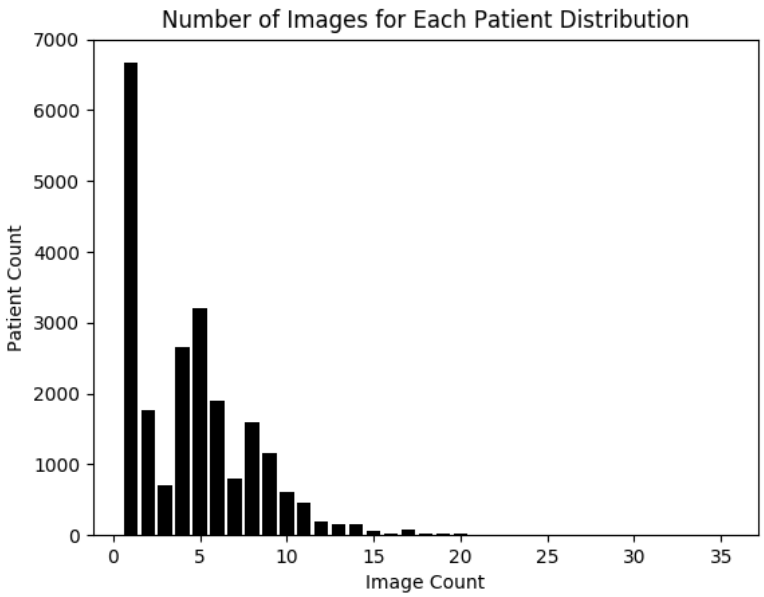


Figure 3.1: Patient Count for Given Number of Mammogram Image

The following table summarizes the distribution of the images across different views. Some of the views have less than 20 images and are not included in the table. Those views include the LTAN(left tangential), RTAN(right tangential), LLMO(left lateromedial oblique), RLMO(right lateromedial oblique), LLM(left late mediolateral), RLM(right late mediolateral), LAT(left MLO with axillary tail modifier), RAT (right MLO view with axillary tail modifier), LCV(left cleavage), RCV(right cleavage). Among the views that are included in the table, LCC(left craniocaudal),RCC (right craniocaudal), LMLO(left mediolateral oblique), RMLO(right mediolateral oblique) are considered standard views required for a mammogram. On the other hand, LXCCL(left exaggerated craniocaudal), RXCCL(right exaggerated craniocaudal), LXCCM(left exaggerated craniocaudal), RXCCM(right exaggerated craniocaudal), LML(left mediolateral), RML(right mediolateral), L(left view position missing), R(right view position missing)) are also mostly supplemental view positions that are non-standard.

Total	CAD	LCC/RCC	LMLO/RMLO	LXCCL/RXCCL	LXCCM/RXCCM	LML/RML	L/R
101494	16417	20255/20353	20273/20288	1465/1645	70/64	165/120	168/163

Table 3.1: number of images by image type

In consideration of the availability of the data and the general requirements of a standard mammogram, we will only be using LCC, RCC(left craniocaudal/right craniocaudal) and LMLO, RMLO(left mediolateral oblique/right mediolateral oblique) view for the experiment. Thus the final total available dataset is comprised of 79501 images. In addition to the image data, there are two types of label that come with the image.

- Exam-level BIRADS level label that indicates the radiologist initial diagnosis result for the patient based on the screening mammogram.
- Pixel-level lesion annotation within the image that indicates the region that the radiologist find suspicious when the diagnosis was done. The reason for the suspicision could be several different lesion patterns such as calcification, masses.

BIRADS(Breast Imaging Reporting and Data System) level labels are used when we make the decision about whether to treat patches from the patient's breast as being normal/negative or lesioned/positive. Specifically, BIRADS level of 1 means that the radiologist believes the patient's breast tissue is completely normal, BIRADS level of 0 means the radiologist found some suspicious pattern within the patient's mammogram, BIRADS level of 2 means the patient has benign cancer. Therefore, for our purpose of the experiment, we will only keep images from patients with these three BIRADS levels in the patch generation process that will be explained in the next section.
The figure-3.2 shows that the majority of the population in UPITT dataset are around the age of 60.
The figure-3.2 shows the BIRADS level distribution across the patients(infrequent BIRADS levels are discarded). Despite of the fact that over 50% of the patients have a BIRADS level of 0 and 2(suspicious and benign cancer), majority of the images/views for benign patients are actually unannotated, thus leading to a scarcity of positive cases.

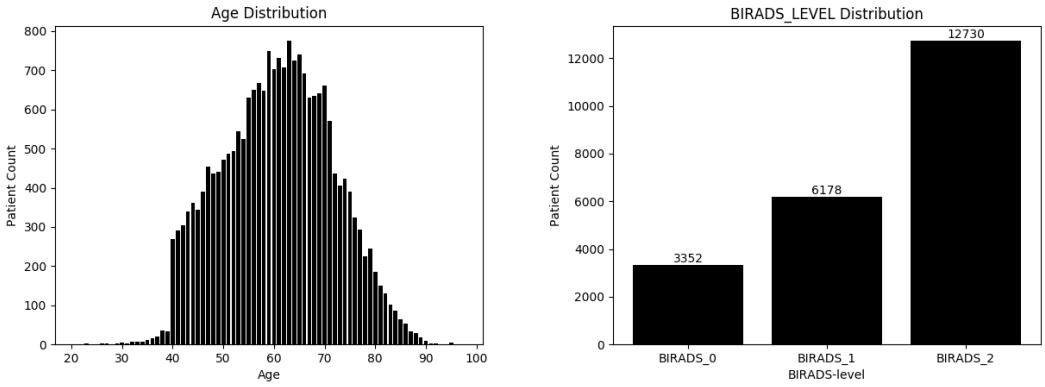


Figure 3.2: Dataset age/BI-RADS level distribution

3.1.2 Dataset preprocessing

Some of the features of a mammogram could significantly impact on the quality of the patch extracted and thus affect the model performance. For the purpose of achieving high consistency in the quality of the patches. The following tables summarizes our criteria used for filtering the images and the the number of images that are discarded based on these criteria. According to table-3.1, we start with **101494** images, after discarding images based on the following standards, we are left with **79501** images for our model development.

Criteria	number of images discarded
Images with ViewPosition other than ”MLO” or ”CC” for the sake of better generalization	20281
Images with SEX that is not ”F” as the data for male is scarce and the diagnosis standards could be drastically different between male and female	9
Images with Breast Implant Present being ”YES” as the presence is rare and the patch for the implant area could become noise during training	1752
Images with PresentationLUTShape being ”INVERSE” as the pixel intensities for such images are inverse of regular images and could confuse the model	2
Images with ImageType containing ”ORIGINAL” as those images went through different processing from other images	5

Table 3.2: number of images filtered by criteria

3.1.3 Patch Extraction Algorithm

We now describe the general algorithm that we use to extract the patches from the **79501** images. The algorithm is configured to ensure that the radiologist's annotated areas are accurately extracted as clean lesion patches. The reason why we rely on using the annotations from the radiologists is because the BIRADS level itself does not really tell us whether there is anything that the radiologist finds suspicious when a diagnosis is made. The white elliptical area annotated by the radiologist actually serve as the ground truth for the patch-based lesion classifier. On the other hand, for the normal patches, we just have to ensure that they are actually breast tissues from the patient that indicate absence of any lesion since any part of the breast could have some form of lesion and an absence of an annotation by radiologist for a specific area within the breast can be considered as a signal that the area is normal. There could be rare cases when radiologist missed the annotation of a lesion area but the model performance shouldn't be significantly impacted.
The following pipeline is designed for the extraction of two types of patches and the details for each step will be explained. The overall patch generation pipeline can be divided into two unrelated procedure, the normal patch extraction process and lesion patch extraction process. The lesion patch extraction involves an additional step of elliptical annotation detection in the image at first. After which a fixed size lesion patch will be center cropped from the annotated ellipse. The whole pipeline is implemented based on the python's cv2, numpy, scipy, pillow, pydicom, pandas package. The plots are created using matplotlib. [13, 16, 17, 23, 26, 37, 39]

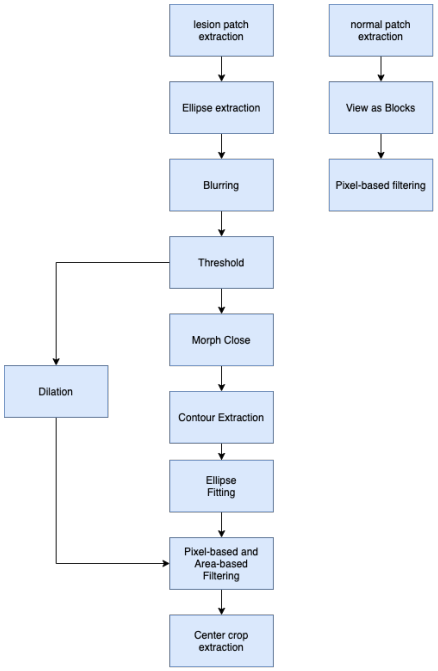


Figure 3.3: Patch Extraction pipelines

Lesion Patch Extraction

1. **Annotated Ellipse Extraction:** Patients with a BIRADS level score of 0 or 2 will undergo the annotated ellipse extraction procedure even though it is not guaranteed that such annotation exists for all patients. Figure 2.2 shows how extracted patch looks like and the form of lesion could be mass, calcification, etc. Each step of the pipeline is explained below with the intermediate results shown for illustration purpose.
2. **Blurring:** Median filtering is applied to the input image for smoothing to improve the accuracy of contour detection. The major advantage of using median filtering is that it preserves the edge while removing the noises from the image.

Figure 3.4: Blurred mammogram using median filtering with radius of 5

3. **Thresholding:** The purpose of thresholding is to bring the outlines of the elliptical annotations by the radiologists to the foreground by using a filter on the RGB pixel intensities.

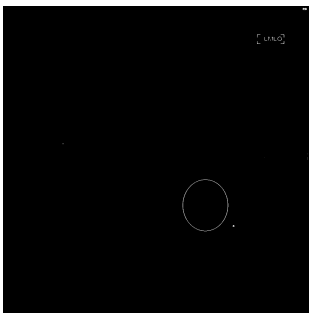


Figure 3.5: Thresholded image based on pixel intensities of annotations

4. **Morph Closing:** Morphological closing is basically a dilation operation followed by erosion operation. Formally,
   erosion of A by B: set of all x s.t. $x + b \in A$ for every $b \in B$

$$A \ominus B = \{x \in E^N | x + b \in A \text{ for every } b \in B\}$$

   dilation of A by B: $A \oplus B$

$$A \oplus B = \{c \in E^N | c = a + b \text{ for some } a \in A \text{ and } b \in B\}$$

   The major purpose of using morphological closing is also to enlarge the boundary of the foreground. Its major difference from dilation is that it is less destructive of the original boundary shape so that the contour fitting will become more accurate. Figure 3.6 shows how the boundary for the annotation thickens and the text boundaries also merged as a block.

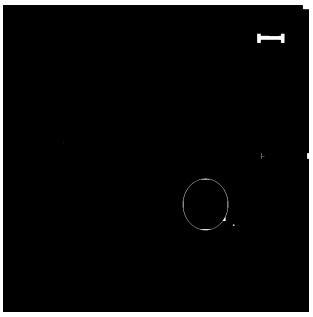


Figure 3.6: Image after morph closing

5. **Dilation:** The edges of the thresholded boundaries are slightly dilated to serve as a mask that will be used later in the ellipse filtering stage.

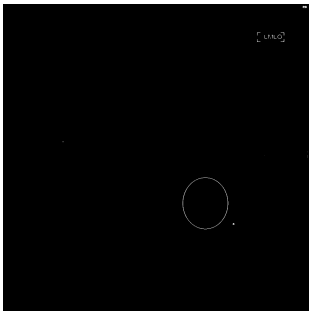


Figure 3.7: Image after dilation of two iterations with raduis of 5

6. **Contour Extraction:** Closed curve detection is performed on the image from step 4. The boundaries obtained from the edges may contain the boundary for the radiologist annotation.
7. **Ellipse Fitting:** Following the contours extracted in the previous step, ellipses will be fit and marked in green as shown in Figure 3.8.

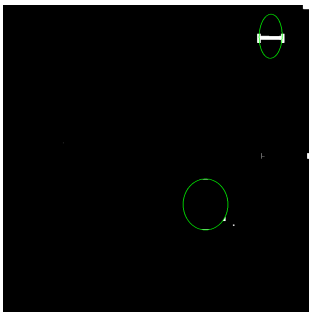


Figure 3.8: Extracted ellipse marked in green

8. **Pixel-based and Area-based filtering:** Note how the text's contour also has an ellipse fitted on it in Figure 3.8. The filtering step uses several metric to ensure that the ellipse extracted is actually the annotation by the radiologist. It first checks whether the contour area and ellipse are is larger than $MIN\_CONTOUR\_AREA$ and $MIN\_ELLIPSE\_AREA$ that are pre-configured based on experiments with a list of patients. Then it uses the previously dilated image as the mask to check the overlapping between the boundaries of the ellipse fitted and the white annotations marked by the radiologist. We only keep the patch if the percentage of the overlapping is greater than $MIN\_ELLIPSE\_COVERAGE$, which is set to 0.999 to ensure that the ellipse fitted is indeed marked by the radiologist and the patch stays within the ellipse. This is how the ellipse fitted on text is removed.
9. **Center crop ROI:** By using the center crop method, we extract the ROI(region of interest). We fit a patch of BLCK_SIZE (128x128) at the center of the fitted ellipse by specifying the top-left corner and bottom-right corner as (center_x $- 0.5 \cdot$ BLCK_SIZE, center_y $- 0.5 \cdot$ BLCK_SIZE) and (center_x $+ 0.5 \cdot$ BLCK_SIZE, center_y $+ 0.5 \cdot$ BLCK_SIZE). We filter the patch based on the coverage of purely white/black pixels and thus ensure that the

patch does not contain the white annotation mark from the radiologist which might distort model predictions.Figure 3.9 shows the patch that is removed based on this criteria.

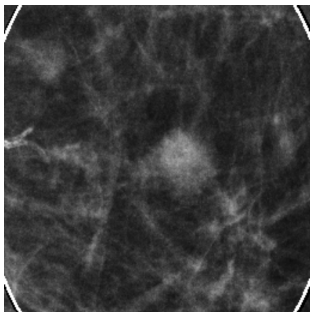


Figure 3.9: A sample patch that is filtered

10. **Output:** Patch containing the ROI (region of interest) that is marked by the radiologist and does not get filtered.

Normal Patch Extraction

1. **View as Patches/Blocks:** The original mammogram is broken into blocks of size $BLCK\_SIZE$ (128x128) with black padding using the scipy view_as_blocks function.
2. **Pixel-based filtering:** Iterate through the blocks, for each block, its quality is first checked based on the distribution of its pixels' RGB intensities. If majority of the pixels' intensities are below a threshold or above a threshold, the patch is considered as being mostly black/white such that it contains no meaningful information and will be dropped. The RGB thresholds for black/white pixels are determined based on the RGB intensities of the pixels in the black background and pixels from the radiologists annotations.
3. **Output:** Patches that contain the breast tissue will be saved as Normal patch data.

Patch statistics

After the dataset pre-processing and patch extraction logic is applied to the dataset, we are left with **976838** normal patches and **3783** lesion patches. The ratio of negative/normal patches to positive/lesion patches is approximately **258:1**, indicating quite serious data imbalance issue for the UPITT-mammogram dataset. The normal patches come from **2952** different patients. The Figure 3.10 summarizes the distribution of the lesion patches across different BI-RADS level. **3737** lesion patches come from **1930** BIRADS_0 patients whom the radiologist find suspicious and need further screening. However, for majority of the benign patients, their mammograms are unannotated, with only **46** patches coming from **32** different BIRADS_2 patients. The imbalance between the number of lesion patches and normal patches is significant for two major reasons.

- The inherent nature of all medical imaging data. Majority of the mammograms actually come from normal patients without benign/malignant cancer or visible lesion that requires further screening.

- The ROI(region of interest) that is suspicious as annotated on the mammogram usually covers a much smaller region than the area that the normal patches could come from, which is essentially the whole breast. So majority of normal patches are fatty tissue despite of the fact that lesions usually occur in glandular, connective tissue.

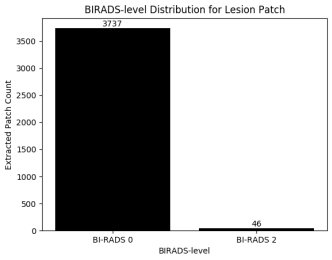


Figure 3.10: Lesion Patch BIRADS-level Distribution

3.2 Data Augmentation

We now describe how we apply deep learning in data augmentation that could potentially help us achieve better generalization performance in the medical image analysis task. Specifically, we will explain two algorithms that we used for generation of new data from the minority lesion class to reduce the huge gap in data availability between normal and lesion class and to alleviate the bad generalization performance issue that is pointed out in the NYU experiment [41]. We will evaluate the data augmentation algorithms based on whether they would improve the state of the art performance that is achieved by our baseline classifier. One of them is based on the traditional SMOTE-based interpolation method and takes a step forward by combining it with a pre-trained feature model and the other approach uses GAN to generate minority class data based on a given high-dimensional latent vector. Different loss functions and regularization techniques were applied to optimize the performance of the GAN. We present the architecture that is used for deep learning and provide training details and evaluate the approaches in next section.

3.2.1 Image Based Embeddings

Multiple studies have shown how important the embeddings are in machine learning tasks[30, 36]. The success of SiameseCBOW showcases that the quality of sentence embeddings can be optimized by averaging the word embeddings generated from a Siamese network.[19] We take the inspiration from SiameseCBOW and SMOTE and attempt to acquire an optimized high-level feature representation for the lesion images using a deep-convolutional architecture before application of SMOTE algorithm based on the distance in the projected feature space. The data augmentation techniques seek to fill in the gaps between data points in high-level feature space such that a more accurate decision boundary can be used to distinguish the data points and reduce over-fitting. So we can essentially summarize our problem as:
Given dataset $D = \{I_1, I_2, \ldots, I_N\}$ images where $I_i \in R^{m \times n \times 1}$ where m, n are the height and width of the grayscale patch images. We want to find a function $\mathcal{F} : R^{m \times n \times 1} \rightarrow R^d$ which maps

the images to a d-dimensional feature vector. Given a loss function $\mathcal{L}$, for simplicity, let it be cross entropy loss where $\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} ylog(\hat{y})$ where y is the true classification label for an image and $\hat{y}$ is the model estimated probability and $\hat{y_i} = M(f_i)$ where $f_i \in R^d = \mathcal{F}\big(I_i\big)$ and M is our normal/lesion patch classifier. Our goal in this medical image analysis is to find the $\mathcal{F}$ that minimizes the loss when we fix the architecture of the classifier M. The major challenge in learning a $\mathcal{F}$ that produces a close to optimal decision boundary that is only susceptible to natural noises is that the training samples that belong to the lesion class is lacking and the gaps between minority samples could potentially lead to a decision boundary that overfits to the training data.

3.2.2 Deep Learning Based SMOTE

Now we describe the deep learning based SMOTE that is inspired by the original smote algorithm in pseudo-code. The interpolation involves the application of a pre-trained model for projection of the image to feature space. The architecture for the feature model will be described later as it is part of the lesion classifier.

Algorithm 1 DL-based SMOTE

Data: $D^m \subset D$: minority samples of the image dataset, k: number of nearrest neighbors, n_m: number of minority samples.
Result: k synthetically interpolated minority feature vectors for each minority sample in the original dataset, so total of $k \times n_m$ samples.
$\mathcal{F}$: pre-trained model/function that maps the image samples from the minority class to d-dimensional embeddings.
dist_func $\in$ {cosine_similarity, $l_1, l_2, \dots$}: a distance metric that measures the similarity between two vectors.
for $i = 1, 2, \dots n_m$ **do**
    Initialize empty max heap denoted as h.
    Apply $\mathcal{F}$ to sample D_i and save f_i
    for $j = 1, 2, \dots, n_m$ **do**
        if i == j **then**
            Skip
        else
            Apply $\mathcal{F}$ to sample D_j and save f_j
            Apply dist_func to f_i and f_j and push to h along with the index j.
            if size(h) $>$ k **then**
                pop the max from h
            end if
        end if
    end for
    for $dist_{\text{neighbor}}, idx_{\text{neighbor}, f_{\text{neighbor}}} \in$ h **do**
        Apply algorithm 2 to f_i and f_{neighbor} and save the returned result.
    end for
end for

Algorithm 2 Interpolation

1: **Data**: f_i: The base feature vector from which synthetic vector is interpolated f_{neighbor}: one of the k-closest neighboring vector.
2: **Result**: A synthetic vector f_{syn} interpolated from f_i to f_{neighbor}
3: Assign $f_{\text{diff}} = f_{\text{neighbor}} - f_i$
4: Generate random float p between 0 and 1.
5: Assign $f_{\text{sync}} = f_i + p \cdot f_{\text{diff}}$ and return

As interpolation in the original pixel space neither produces a valid minority class image nor carries any real significance in the high-dimensional feature space, we would like to take advantage of deep learning and first project the images into high-dimensional feature space and carry out the interpolation there in the hope of achieving the ideal state indicated in figure 3.11.
The question is at what level in the feature extraction process could we apply the interpolation to obtain the neighboring data that could actually help with the training process and generalization. Furthermore, which metric should we apply to quantity the distance between the data points in the feature space. Essentially, we have to use the configurable K-nearest neighbor algorithm[10] to select the k closest neighbors for each real data sample in our dataset based on a specific distance metric. The potential distance metrics that could be used are l1, l2 distance, cosine similarity metric.
After we select the k-nearest neighbor based on a given distance metric, a randomly scaled interpolation is carried out to create a synthetic data sample. By the initial hypothesis of SMOTE, this could potentially help us fill in the gaps between the original minority training data points and achieve better generalization performance in the end.

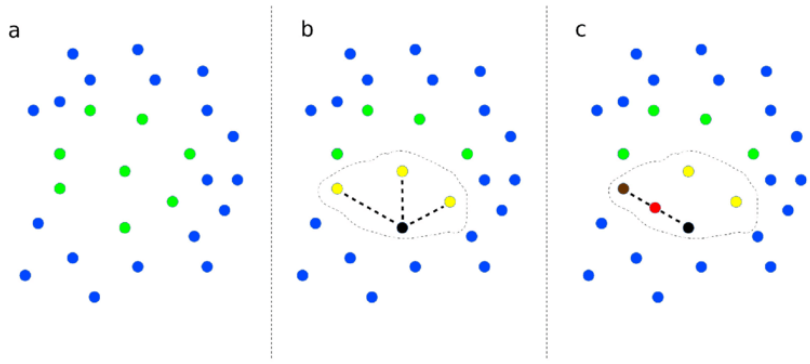


Figure 3.11: SMOTE Mechanism Illustration from [31]

3.2.3 GAN-based Augmentation

Now we describe the procedure for generation of data samples using a GAN trained specifically for data augmentation. Our hypothesis is that the GAN would add data points to our original training dataset to help fill in the gaps by projection from an arbitrary noise vector at the cost of

adding some noises to the original dataset. As the augmentation is only applied to the minority class, our goal is to improve the recall while minimizing the decline in the precision as it is quite inevitable that there exists such a trade-off between recall and precision when we fix the model architecture.Since the model is exposed to more randomly generated synthetic data, noises from the GAN could actually mislead the model to predict more false positives. However, if the GAN is really effective in producing high-quality synthetic data, we could potentially improve the final F1-score of our model by significantly improving the recall without losing too much precision.

GAN

The GAN(generative adverserial network) applied in our experiment with the UPITT-mammogram dataset is based on a deep-convolutional architecture. The architecture is summarized in Figure 3.12. The discriminator and generator basically have symmetric architectures such that one down-samples the image to output its conviction about whether the image is fake or real and the other up-samples a fixed latent vector to an image.

In this experiment, we take the WGAN[6, 12] as the baseline and compare its performance against relavistic-GAN[18] with/without spectral normalization[24] and TTUR(using different learning rate for discriminator and generator) for stabilizing the training process. Similar to gradient penalty, the spectral normalization achieves improved stability by enforcing the lipshitz constant and the TTUR is applied to ensure tha the discriminator does not get too powerful before the generator learns, thus avoiding the gradient vanishing or gradient explosion that stalls the learning process too early.

The relavistic-GAN applies the discriminator to both the real and fake images while training both the discriminator and the generator while a traditional WGAN[6, 12] does not really feed the real images to the discriminator while training the generator. The loss function seeks to minimize the difference in the confidence score from the discriminator instead of just trying to maximizing the confidence score for fake images. The following loss functions illustrate this fundamental difference. (D represents the discriminator, P is the real distribution of minority class data, Q is the modeled distribution, f_1, f_2, g_1, g_2 are scalar functions)

$$L_{discriminator} = -E_{x_r \sim P}[f_1(D(x_r))] + E_{x_f \sim Q}[f_2(D(x_f))]$$
$$L_{relavistic\_discriminator} = E_{(x_r,x_f)\sim(P,Q)}[f_1(D(x_r) - D(x_f))] + E_{(x_r,x_f)\sim(P,Q)}[f_2(D(x_f) - D(x_r))]$$
$$L_{generator} = -E_{x_r \sim P}[g_1(D(x_r))] + E_{x_f \sim Q}[g_2(D(x_f))]$$
$$L_{relavistic\_generator} = E_{(x_r,x_f)\sim(P,Q)}[g_1(D(x_r) - D(x_f))] + E_{(x_r,x_f)\sim(P,Q)}[g_2(D(x_f) - D(x_r))]$$

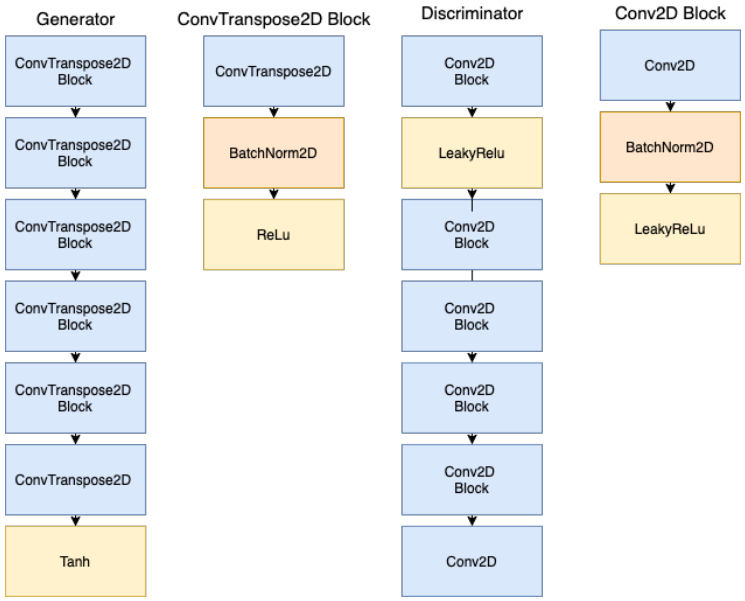


Figure 3.12: GAN Architecture

The following code snippet describes how we generate synthetic data by applying the pre-trained generator from the GAN to the noise vector in the latent space. In this experiment, we generate 7160 samples that essentially triple the total number of lesion patches. In the evaluation section, we will explain how we evaluate the GANs and choose the best performing GAN for the generation of the synthetic data.

Algorithm 3 GAN-Based Augmentation

1: **Data**: n_m: number of minority samples.
2: **Result**: n_m synthetic minority class sample
3: $\mathcal{G}$: pre-trained generator in GAN that takes in a noise vector and generate minority class image.
4: **for** $i = 1, 2, \ldots n_m$ **do**
5: Sample a random noise vector in the latent space and save as noise_vec.
6: Apply the generator to noise_vec and save the result $\mathcal{G}(noise\_vec)$
7: **end for**

3.2.4 Lesion Classifier

Now we will describe the classifier that we applied for the original training dataset and GAN augmented dataset. As the training pipeline for the original dataset, GAN-augmented dataset is different from the SMOTE-augmented dataset. We will describe the two pipelines separately.
The basic classifier architecture we used in this experiment is similar to VGG.[35] The reason why we used this architecture instead of Resnet is because based on our early experiments, despite of the fact that the VGG model has lower complexity, the model's generalization ability

is not significantly different from that of a model based on Resnet18 architecture. However, the convergence speed is faster because of lowered model complexity.

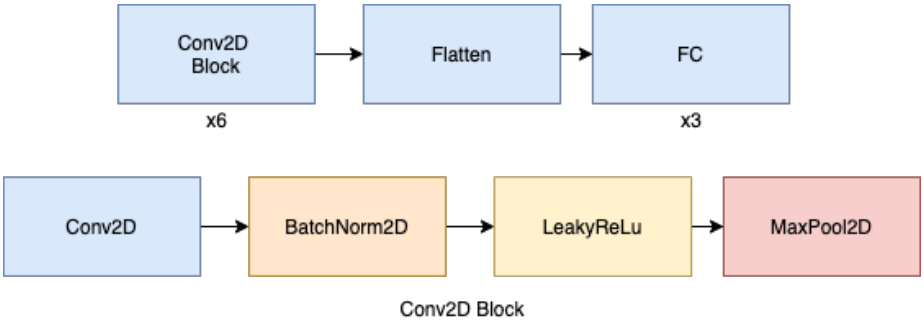


Figure 3.13: Classifier Architectures

For SMOTE augmentation, we split the training process into three different phases.

The first phase, pre-training phase, is also based on the same architecture in Figure 3.13, we train the model for the same number of iterations and save the weights for the first convolutional block and the rest of the model separately. The decision is made to test our hypothesis that interpolation in the low-level feature representation for the images would more likely hit feature representation for unseen data samples than interpolating from the data samples directly. Further experimentation is necessary to evaluate the depth at which we apply the interpolation though.

In the second phase, interpolation phase, we utilize the weights from the first convolutional block and basically follows the Algorithm 1 to create and save the synthetic feature vectors, in this experiment we set k = 2 so we essentially triple the set of lesion samples by interpolating 2 synthetic data samples from each original data sample.

In the third phase, we use the pre-trained weight from the first convolutional and set its requires_grad parameter to False to pause its learning. We concatenate the feature vectors saved with the embeddings for training images before feeding them into the rest of the architecture.

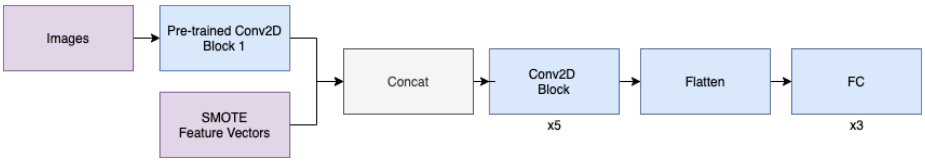


Figure 3.14: SMOTE Classifier Architectures

In order to obtain a relatively accurate estimation of the final model's generalization performance, we split all the patches into training, validation and test set by a 50%, 25%, 25% ratio. A large portion of the dataset is dedicated for testing to acquire an accurate reflection of the generalization ability of the model. The validation performance provides us with a gauge on how the model may preform on the test set. The split is done randomly based on patients so we ensure

that patches from the same patient will not show up in both the training and test set to avoid the problem that the model sees a patch in the test set from a patient whose other patches showed up in training, which could cause data leaking and distort model performance. The seed for random splitting, sampling, shuffling process is fixed to ensure the training and test dataset is fixed across the experiments to make the results replicated.

Chapter 4

Evaluation

In this section, we will evaluate the performance of the deep learning models that we used for dealing with the severe data imbalance issue in the UPITT-mammogram experiment with the syntheic data from deep learning based SMOTE and GAN on classifying the normal/lesion patches. The highlights of evaluation includes

- Achievement of an FID score of **58.5** by using WGAN-gp(64 filter in the first layer) with spectral normalization [6] and TTUR. The two regularization techniques, spectral normalization [24] and TTUR, significantly stabilized model performance.
- The GAN based data augmentation technique improves the overall f1-score for view-wise and patient-wise classification by **2.5**% and **1.4**%, the recall is quite significantly improved by **5.3**%, **5.3**%, **4.3**% respectively at the cost of minor decreases in the precision.
- The DL-based SMOTE data augmentation technique that uses cosine_similarity distance function improves the overall f1-score for view-wise and patient-wise classification by **0.4**% and **0.7**%, the recall for patch-wise classification is improved by 0.2%. Furthermore, the precision is also slightly improved by **0.9**%, **1.5**% for view-wise and patient-wise classification.
- The DL-based SMOTE data augmentation technique that uses l1 distance function improves the overall f1-score for patch-wise, patient-wise classification by **0.1**% and **0.4**%, The precision is slightly improved by **0.3**%, **0.6**% for view-wise and patient-wise classification.
- The DL-based SMOTE data augmentation technique that uses l2 distance function improves the overall f1-score for patch-wise, view-wise and patient-wise classification by **0.3**%, **0.3**% and **0.6**%, the recall is slightly improved by **0.4**%, **0.1**%, **0.2**% respectively. Furthermore, the precision is also slightly by **0.1**%, **0.5**%, **0.8**% along with improved recall.

4.1 Evaluation setup

The implementation is based on pytorch, sklearn, numpy.[13, 27, 39]

Dataset. We evaluate our augmentation approaches with the UPITT-mammogram dataset, statistics of which is documented in 3.1. We group the patches extracted from the mammograms based on the patient id and obtain two lists of patients, differentiated based on their BIRADS

level score. We select a fixed random seed and split the patients for each list based on a 50:25:25 for training, validation and testing and use the validation performance to save the model parameters to be evaluated for testing. We repeat this process for multiple times to ensure that the test performance collected is not randomly favoring one method over another.

Training configs. The following tables describe the hyperparameter configurations that we used for training the GAN.

Params	Value
Block1 filter	64/128
Latent size	64/128
Discriminator Learning rate	2e-4
Generator Learning rate	2e-4
Batch Size	64
Adams Optimizer Beta1,Beta2	0.5,0.999

Table 4.1: GAN non-TTUR training hyperparameters

Params	Value
Discriminator Learning rate	4e-4
Generator Learning rate	1e-4
Adams Optimizer Beta1,Beta2	0.0,0.9

Table 4.2: GAN TTUR training hyperparameters

Params	Value
Block1 filter	16
Learning rate	1e-5
Dropout	0.5
Batch Size	128
Adams Optimizer Beta1,Beta2	0.5,0.999

Table 4.3: Classifier training hyperparameters

Training environment. We use the AWS p3.2xlarge instance equipped with one Tesla-V100 GPU with 16 gigabytes of memory. The instance has 8 vCPUs, 64GB memory and up to 10 Gbps network bandwidth.

Synthetic data evaluation metric There are two common objective metrics that could be applied for an objective evaluation of the quality of the synthetic data produced by a GAN. IS(Inception Score) and FID(Frechet Inception Distance), both of which are based on a pre-trained inception V3 model. The t-SNE[38] technique could be used for visualization of synthetic data features.

- The inception score basically estimates the quality of the synthetic images compared with the 1000 objects within its training dataset from ImageNet. As lesioned breast tissue patch is not among the objects in the ImageNet, so the inception score is relatively low regardless of the actual synthetic patch quality, which is the reason why we only consider it as a reference but not criteria for the selection of the final model for synthetic data generation procedure.
- The FID is the distance between the gaussian computed mean and variance based on the activation for the batch of synthetic images as opposed to the batch of real images. Therefore, for our use case, the FID is the more reasonable measure for the synthetic image quality for final model selection.
- The t-SNE[38] is a common dimensionality reduction technique that allows us to project the high-dimensional feature vectors acquired from a pre-trained feature model to lower dimensional space. This method allows us to subjectively evaluate the synthetic data produced. However, distance in 2-dimensional space is not directly proportional to the distance in high-dimensional space as t-SNE[38] does not really retain the distance in high-dimensional space during the reduction process via gradient descent.

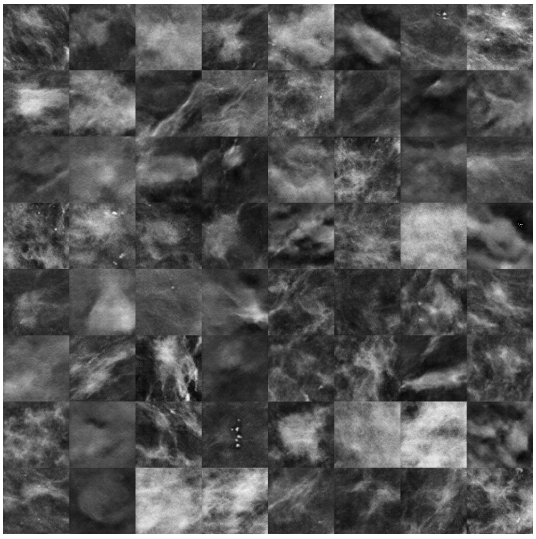
Classifier metrics for evaluation. There are several different ways to evaluate the quality of the patch-based lesion classifier. The classifier could also have quite different performance when evaluated using different metric.

- Accuracy: The simplest metric that is computed using $\frac{TP+TN}{TP+TN+FP+FN}$. However, the simple metric is not that suitable for tasks that have really imbalanced dataset as the correctness of negative/majority class dominates the final result and a naive classifier could achieve meaninglessly high accuracy.
- Patch-based Precision, Recall, F1 score and MCC. Precision = $\frac{TP}{TP+FP}$, Recall = $\frac{TP}{TP+FN}$, F1 score = $\frac{2*TP*TP}{2*TP+FP+FN}$, MCC = $\frac{TP*TN-FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}$. These three metrics put more emphasis on the correctness of the positive/minority class. A naive classifier would have really low F1 score if it simply predicts everything as negative. However, the flaw for relying on F1-score is that it is still susceptible to minor shifts in the prediction for negative class because increase in false positive rate could significantly lower F1-score as the number of negative samples is large. This is the reason why we also include MCC as it includes all the values from the confusion matrix and is more suitable for imbalanced dataset by including TN.
- Mammogram image-wise and patient-wise Precision, Recall, F1 score and MCC. The reason why we incorporate the view-wise and patient-wise metrics is because we want to have a more accurate reflection of how the classifier could actually be performing in reality and we want to reduce the penalty for slightly loosening the threshold for predicting a patch as

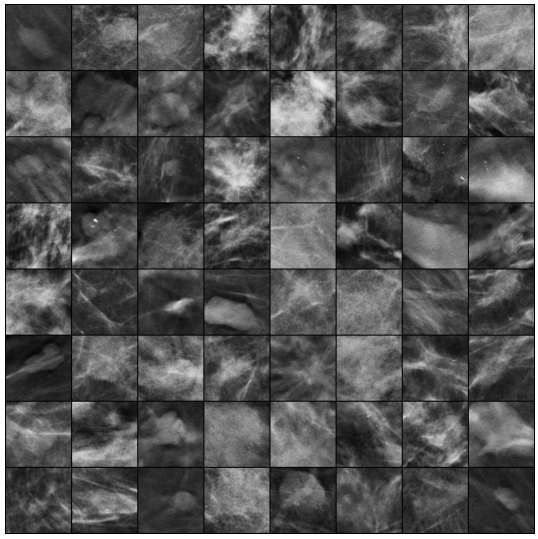
positive. A mammogram image is classified as normal only if all the patches within it are predicted as normal. A patient is predicted as normal only if all views within the patient folder is predicted as normal.

4.2 GAN performance evaluation

We first offer a comparison of the random synthetic lesion samples from GAN and the actual lesion patches. The whole training process takes 300 epochs. We save the model parameters when the lowest FID score is achieved. By comparing the images with 2.2, we could notice how the generator has picked up certain patterns such as calcification and mass and most of the samples look quite realistic. The synthetic data includes more variations of those forms of lesion at the cost of some lower quality noisy samples. The question is whether these noises would significantly hurt the model's precision when we try to improve the recall by inclusion of more minority class samples.



(a) WGAN-gp_spec random lesion samples



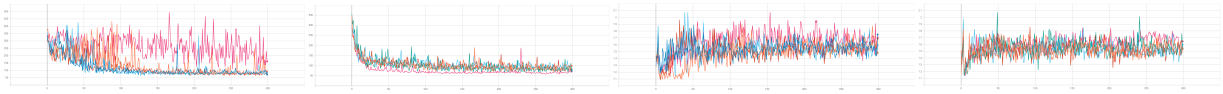
(b) Real random lesion samples

Figure 4.1: GAN synthetic image quality comparison

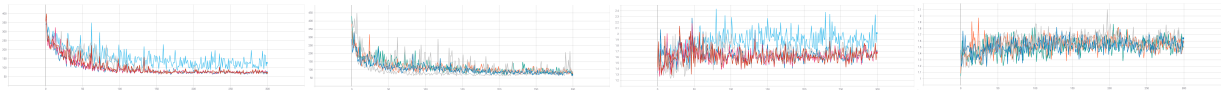
Now we evaluate the impact that different loss functions have on the final performance in terms of the FID score and IS score. First, we show the impact of spectral normalization and TTUR and number of filters on performance and convergence.

metric/spec+TTUR	rahinge	ralsgan	rasgan	rsgan	wgan-gp
FID(64 filter)	66.9/68.9	66.2/65.8	**64.1**/71.9	66.2/71.0	99.8/**58.5**
FID(128 filter)	64.6/**64.6**	**61.9**/67.7	65.6/71.7	**63.0**/69.8	85.6/62.7
IS(64 filter)	1.89/1.82	1.9/2.08	2.08/1.98	1.96/1.86	2.07/1.90
IS(128 filter)	**2.14**/1.82	2.04/**2.1**	**2.18**/1.97	**2.28**/1.87	**2.43**/1.88

Table 4.4: GAN FID and IS



(a) 64-filter FID (b) 64 filter spectral norm FID (c) 64 filter IS (d) 64 filter spectral norm IS



(e) 128 filter FID (f) 128 filter spectral norm FID (g) 128 filter IS (h) 128 filter spectral norm IS

Figure 4.2: GAN FID and IS score

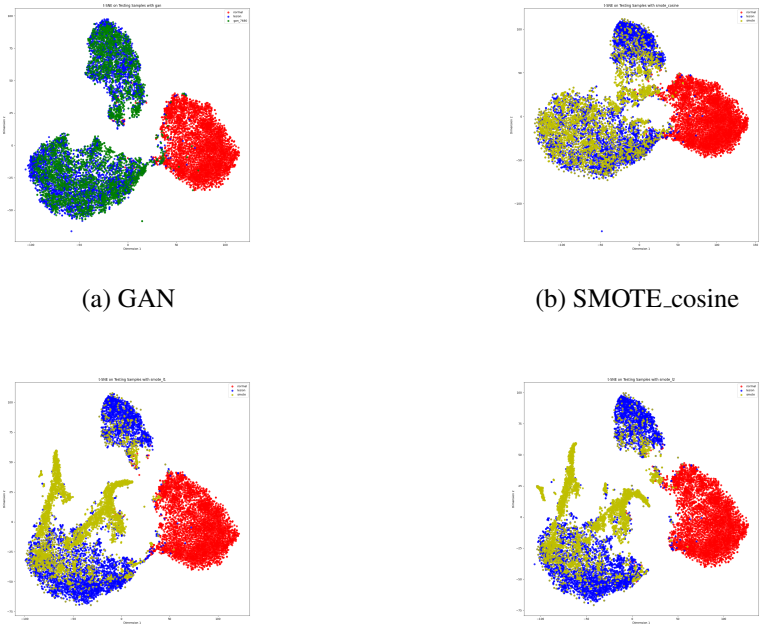
Effect of model complexity/number of filters. Figure 4.2 and Table 4.4 shows the impact of the number of filters on the convergence behavior and the final model performance. When we increased the number of filters to 128(in the first conv layer, and doubling the number of filters for the remaining conv layers), the model's training stability significantly improved. Across the 5 different loss functions that we applied to evaluate the impact of the filter increase, the FID decreased by 2.2 on average and the IS score increased by 0.12, indicating the potential for getting improved performance when the model complexity increases. This is consistent with the conclusions drawn in previous studies such as BIGGAN.[8] We believe that a bigger model could potentially produce more realistic images at higher reolution even when the amount of training data is scarce.

Effect of spectral normalization and TTUR Figure 4.2 shows that the spectral normalization[24] and TTUR could have a quite significant impact on the model stability. The two regularization techniques, by enforcing the lipschitz constraint for the discriminator to be less than 1 and using a lower discriminator learning rate, ensures that the discriminator does not get too powerful such that the gradient vanishes before the generator learns. However, from Table 4.4, we also did not observe improvement in performance mentioned in previous literature despite of the fact that WGAN-gp achieved much better performance with further regularization on top of gradient-penalty. On average, the FID score increased by 3.13 and the IS score decreased by 0.16 with strict regularization.

4.3 Synthetic data t-SNE Visualization

Now we visualize the feature vectors produced by feeding the synthetic data points to the classifier. t-SNE serves as a preview that indicates how the synthetic data could potentially impact on the final classifier performance. In the t-SNE plots in Figure **??**, the **blue**, **red**, **green**, and **yellow** dots represent the projected feature vectors(dimensionality reduced) from the original lesion patch, normal patch, GAN and SMOTE synthetic data respectively. From the t-SNE plots, we could derive several observations

- The synthetic data points are occupying regions that are not covered by the original data points. This indicates that the synthetic data points could possess features that are not covered by the original data points.
- The GAN-based and smote_cosine-based synthetic data points are less noisy and lying in closer distance to the original data points compared with the other two distance metrics in deep learning based SMOTE. The closeness after dimensionality reduction does not necessarily imply closeness in high-dimensional space, though.
- The SMOTE based on the l1, l2 distance function are lying farther away from the original data points. This indicates the potential ineffectiveness of applying l1,l2 when the feature vectors have high dimensions. The SMOTE synthetic data points from l1,l2 significantly expands the original minority class region. We will later evaluate how this translate into final classifier performance.



(a) GAN

(b) SMOTE_cosine

(c) SMOTE_l1

(d) SMOTE_l2

Figure 4.3: t-SNE visualization, blue: lesion, red: normal, green: GAN synthetic lesion, yellow: SMOTE synthetic lesion

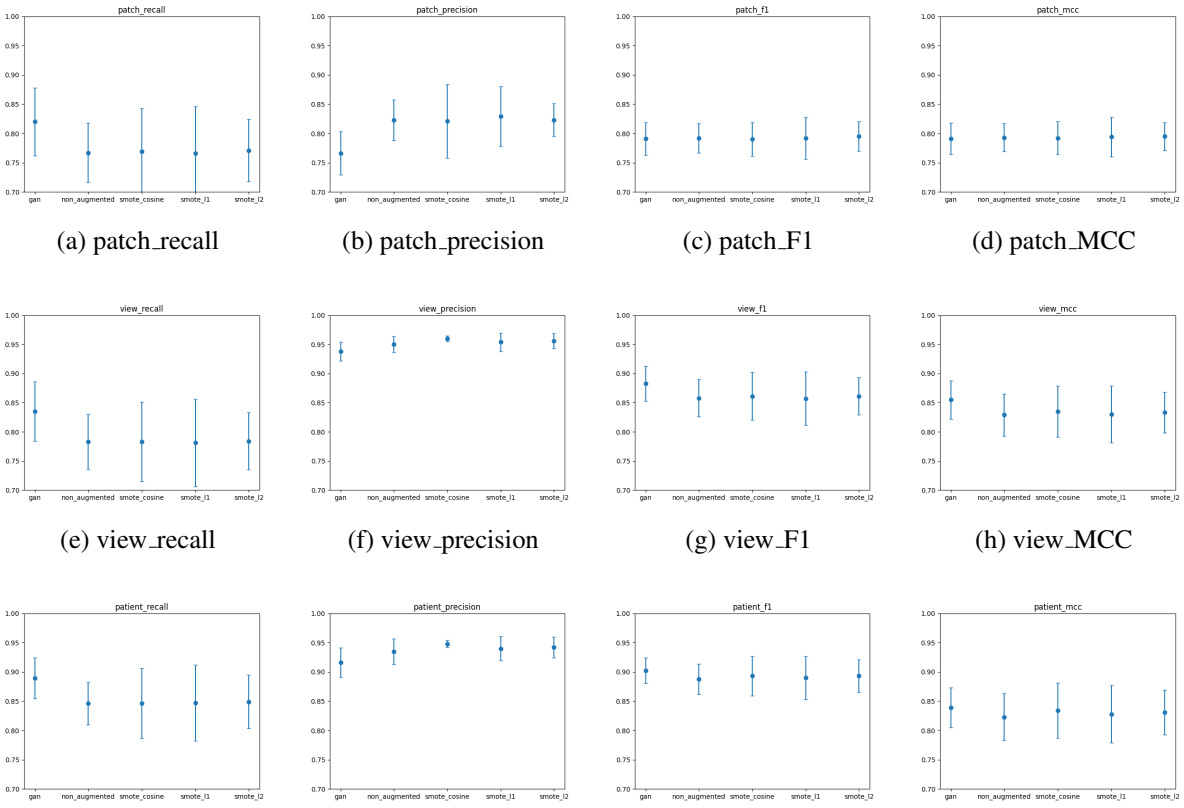
4.4 Classifier performance evaluation

metric	gan	unaugmented	smote_cosine	smote_l1	smote_l2
Patch_recall	**0.82**0	0.767	0.769	0.766	0.771
Patch_precision	0.763	0.821	0.814	**0.823**	0.822
Patch_MCC	0.790	0.793	0.790	0.793	**0.795**
Patch_F1	0.791	0.793	0.791	0.794	**0.796**
View_recall	**0.836**	0.783	0.783	0.781	0.784
View_precision	0.938	0.951	**0.960**	0.954	0.956
View_MCC	**0.856**	0.830	0.835	0.830	0.834
View_F1	**0.884**	0.859	0.863	0.859	0.862
Patient_recall	**0.890**	0.847	0.847	0.847	0.849
Patient_precision	0.916	0.935	**0.950**	0.941	0.943
Patient_MCC	**0.840**	0.823	0.835	0.828	0.832
Patient_F1	**0.902**	0.888	0.895	0.892	0.894

Table 4.5: Classifier test performance summary

We first summarize the average test performance across the five trials that split the train/test dataset differently and then we evaluate the consistency of the results across different trials. From Table 4.5, we could make the following observations

- We observe performance improvements across all metrics with different augmentation techniques. Notably, the F1-score for patch-wise, view-wise and patient-wise classification improved by **0.3**%, **2.5**% and **1.4**%, the MCC, the metric that considers true negative, improved by **0.2**%, **2.6**%, **1.7**% across the 3 levels..
- There is no significant performance difference among the three deep learning based SMOTE augmentation techniques despite of the fact that the SMOTE synthetic data that relies on l1,l2 distance metric are more noisy in the t-SNE plots. This implies that distance after dimensionality reduction is not necessarily proportional to the actual distance in high dimensional space. Overall, we observe a slight improvement, around 0.5% in f1-score and MCC across different levels of classification for SMOTE based augmentation compared with the model performance without augmentation applied.
- The GAN-based augmentation approach achieves our major goal of improving the recall(reducing the number of false negative). By using synthetic data from GAN, the recall improved by **5.3**%, **5.3**%, **4.3**% respectively for patch-wise, view-wise, patient-wise classification. Furthermore, the model almost performs the best across all metrics for view-wise and patient-wise classification.



(a) patch_recall (b) patch_precision (c) patch_F1 (d) patch_MCC

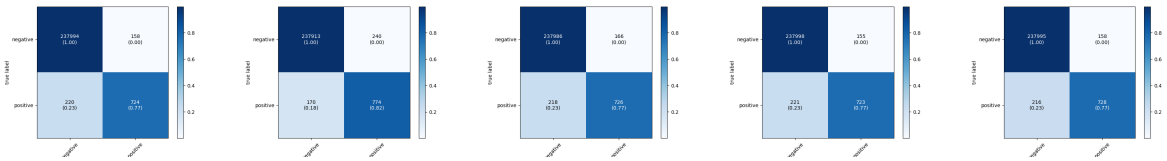
(e) view_recall (f) view_precision (g) view_F1 (h) view_MCC

(i) patient_recall (j) patient_precision (k) patient_F1 (l) patient_MCC

Figure 4.4: Error bar plots for classifier performance

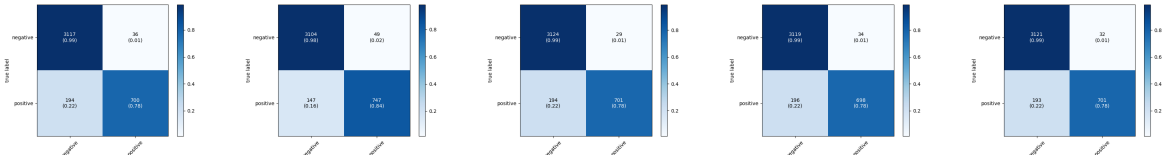
After we have evaluated the models' average performance across different trials, We now evaluate the models' consistency and stability across the trails using an error bar plot.

- Overall, models augmented with GAN synthetic data have similar stability as the unaugmented model. We attribute this to the fact that GAN's synthetic data is created by a generator that tries to fool the discriminator, so the overall feature pattern would be less noisy since otherwise the discriminator could easily distinguish the fake data from real data.
- On the other hand, models augmented with SMOTE based synthetic data has slightly worse stability as the longer error bar indicates more variability across the experiments. Compared with GAN's synthetic data, SMOTE's synthetic data is based on interpolation after first layer of feature extraction and it involves more randomness. It also does not consider whether the synthetic data would be lying close to real data in the higher dimensional space after passing through further feature extractions. As a result, the synthetic data could either cover unseen features and improve the minority class boundary or become too noisy and worsen the model performance.



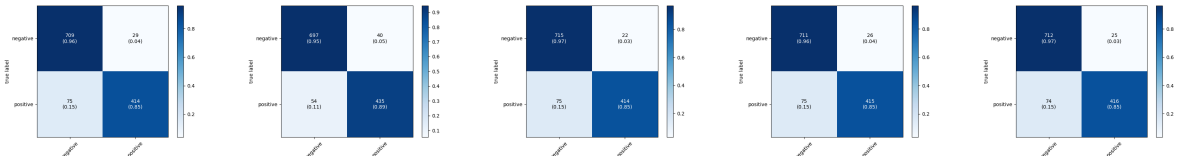
(a) unaugmented (b) gan (c) smote_cosine (d) smote_l1 (e) smote_l2

Figure 4.5: patch-based confusion matrices



(a) unaugmented (b) gan (c) smote_cosine (d) smote_l1 (e) smote_l2

Figure 4.6: view-based confusion matrices



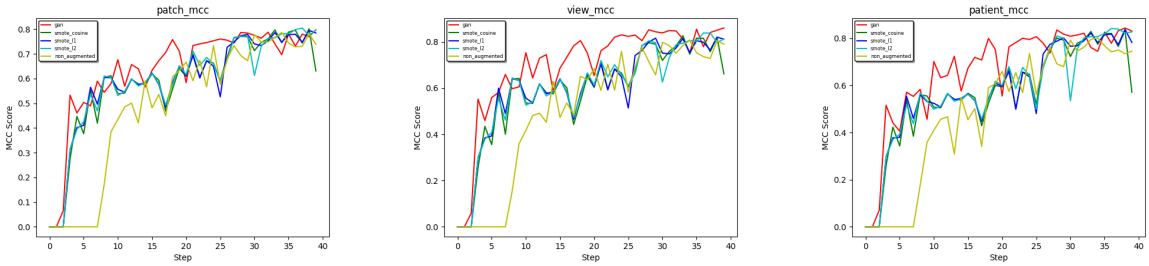
(a) unaugmented (b) gan (c) smote_cosine (d) smote_l1 (e) smote_l2

Figure 4.7: patient-based confusion matrices

Figure 4.5, Figure 4.6 and Figure 4.8 are the averaged confusion matrices across the five different trials. They provide a more straightforward overview of how the models are performing after the augmentation with synthetic data. There are several key observations.

- GAN augmentation decreased the number of false negative across all 3 levels of specificity. This comes at the cost of slightly larger number of false positives. Since the total number of negative cases is large, a relatively small change in precision could have led to this consequence. Overall, our goal is to maximize recall while possibly maintaining the precision.
- For models trained with SMOTE-augmented data, the final prediction results are quite similar. The models' recall does not change significantly from the unaugmented model, but we could still observe that the precision is slightly improved with a more precise decision boundary, but overall the change in performance with SMOTE augmentation is less drastic than GAN augmentation.

Lastly, we look at the MCC plots for five models trained with different augmentation techniques to briefly analyze how does improving the dataset balance affect the model performance. Based on the figures, we derive the following observations



(a) Patch_MCC (b) View_MCC (c) Patient_MCC

Figure 4.8: Sample MCC Plots for Models under Different Augmentation Schemes

- Overall, we could observe that the GAN and SMOTE augmentation slightly improves the convergence speed and the performance plateaus around **20-30** steps.
- There are some instabilities in performance during training. The smaller the number of test samples, the greater the instability, meaning that patch-wise performance is more stable than view-wise performance than patient-wise performance.
- The performance is consistent with the confusion matrices and table summary. We see some small improvements in performance with GAN augmentation and the improvements from SMOTE augmentation are much more subtle and harder to tell from the plots.

Chapter 5

Conclusion

Based on our experiment with the UPITT-mammogram dataset, we have acquired quite solid empirical evidence that deep learning based data augmentation could improve the performance of models trained on an un-augmented or image-manipulation(flipping, rotation) augmented dataset. Data augmentation is a critical technique in addressing the data scarcity and imbalance issues in the field of deep learning based medical image analysis. Compared with traditional image-manipulation based augmentation techniques, the deep learning based SMOTE and GAN push the MCC and F1-score at patient-level up by 0.5-1%/2-4% with the same original training dataset, model architecture and set of hyperparameters respectively. More importantly, in the medical imaging field, we believe that minimizing the number of false negative is more important than reducing false positive, augmenting the minority/positive class effectively helps us improve the recall. While applying deep learning for data augmentation certainly increases the complexity of the solution development and training time, GAN and SMOTE have effectively shown their impacts on the model performance.

Limitations Now we point out several intrinsic limitations involved in the experiment evaluation and our dataset in general.

- The evaluation of the view-wise and patient-wise classifier performance is still flawed. Given a normal patient/view, we can judge the correctness of model by checking whether any lesion prediction is made across all the patches of the breast. However, given a lesioned view/patient, it is hard to objectively judge if the prediction is completely correct. The model could be judged based on whether it makes the same prediction for the annotated patch by radiologists, but we also could not tell if the model is wrong when it predicts an unannotated patch from a lesioned view as positive.
- The evaluation of SMOTE-based synthetic data quality is also difficult. First of all, the SMOTE-based synthetic data is encoded as feature vectors. This means that the only way for us to directly measure its quality is based on the final classifier performance. However, even if the classifier generalization performance is improved after the addition of SMOTE-based synthetic, we could not easily learn about the patterns that the model has picked up from those feature vectors or explain the causes for the change in performance.
- Even though the GAN does produce actual synthetic lesion patches, the evaluation based on

FID and classifier performance changes is still insufficient. Ideally, a group of radiologists should be invited to play the role of discriminator in GAN and evaluate the number of patches they find suspicious patterns among the normal patches, real lesion patches and GAN synthetic lesion patches. The average performance across the radiologists could provide us with a subjective overview of how realistic the synthetic images are and the patterns they have picked up.

- Distance metric for high-dimensional feature vector remains as a research problem. Based on this experiment, we realized that the common l1,l2, cosine_similarity distance applied during interpolation did not result in significant classifier performance difference. A more stable distance metric that is not overly sensitive to one dimension could potentially provide a more accurate reflection of the neighboring relationship among the data points.
- The dataset also imposes several serious limitations in the experiment. The lesion patches are extracted from the annotated images and this impose several limitations. There exists a tradeoff between the number of lesion patches and the resolution/size of the patch. With a higher resolution, we are guaranteed to include the lesion pattern in the extracted patch but we would also be left with a much smaller number of lesion samples as we cannot keep the annotation marks from the radiologists.
- Some mistakenly labeled data also exist in our dataset. We realized that for some of the patients with BIRADS level of 0/2, her mammograms could contain no annotations at all and few patients who have BIRADS level of 1, actually have annotations in the image.

Future Works We propose that future efforts should focus on the following areas

- Evaluation of the classifier's lesion patch performance based on the number of lesion patches that it predict for a given view/patient.
- Design and evaluate other distance metric that are more suitable for high-dimensional vectors. Evaluation of the impact of clipping each individual dimension shift in l1,l2 distance computation on t-SNE visualization and final classifier performance.
- Evaluation of the quality of all patches, including real and synthetic patches, by a group of radiologists.
- Development of more detailed explanation for interpretation on top of annotation explanation that could be provided by a patched-based classifier.
- Development of stronger GAN with larger number of filters and batch_size. (According to the research results on BigGAN by Google Deepmind[8])

Bibliography

[1] Breast cancer, Dec 2020. URL https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470. 1

[2] Diagnostic mammogram, Oct 2020. URL https://www.nationalbreastcancer.org/diagnostic-mammogram. 1

[3] U.s. breast cancer statistics, Feb 2021. URL https://www.breastcancer.org/symptoms/understand_bc/statistics#:~:text=About1in8U.S.,(insitu)breastcancer. 1

[4] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *Lecture Notes in Computer Science*, pages 420–434. Springer, 2001. 2.2.3

[5] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks, 2018. 1

[6] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017. (document), 1, 3.2.3, 4

[7] Eric A. Berns, R. Edward Hendrick, Mariana Solari, Lora Barke, Denise Reddy, Judith Wolfman, Lewis Segal, Patricia Deleon, Stefanie Benjamin, Laura Willis, and et al. Digital and Screen-Film Mammography: Comparison of Image Acquisition and Interpretation Times. *American Journal of Roentgenology*, 187(1):38–41, 2006. doi: 10.2214/ajr.05.1397. 1

[8] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis, 2019. 4.2, 5

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi: 10.1613/jair.953. (document), 1, 2.2.2

[10] Padraig Cunningham and Sarah Jane Delany. k-nearest neighbour classifiers: 2nd edition (with python examples). *CoRR*, abs/2004.04523, 2020. URL https://arxiv.org/abs/2004.04523. 3.2.2

[11] Fabio Henrique Kiyoiti dos Santos Tanaka and Claus Aranha. Data augmentation using gans. *CoRR*, abs/1904.09135, 2019. URL http://arxiv.org/abs/1904.09135. (document), 2.5, 2.2.3

[12] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/892c3b1c6dccd52936e27cbd0ff683d6-Paper.pdf. (document), 1, 3.2.3

[13] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2. 3.1.3, 4.1

[14] Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328, 2008. doi: 10.1109/IJCNN.2008.4633969. 2.2.2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. doi: 10.1109/cvpr.2016.90. 2.1.1

[16] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55. 3.1.3

[17] Itseez. Open source computer vision library. https://github.com/itseez/opencv, 2015. 3.1.3

[18] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard GAN. *CoRR*, abs/1807.00734, 2018. URL http://arxiv.org/abs/1807.00734. (document), 3.2.3

[19] Tom Kenter, Alexey Borisov, and Maarten de Rijke. Siamese cbow: Optimizing word embeddings for sentence representations, 2016. 3.2.1

[20] Sotiris Kotsiantis, D. Kanellopoulos, and P. Pintelas. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30:25–36, 11 2005. (document), 2.3, 2.2.1, 2.2.1

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks. 2.1.1

[22] Yi-Hsun Liu, Chien-Liang Liu, and Shin-Mu Tseng. Deep discriminative features learning and sampling for imbalanced data problem. pages 1146–1151, 11 2018. doi: 10.1109/ICDM.2018.00150. (document), 2.2.3, 2.4

[23] Darcy Mason. Su-e-t-33: pydicom: an open source dicom library. *Medical Physics*, 38 (6Part10):3493–3493, 2011. 3.1.3

[24] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *CoRR*, abs/1802.05957, 2018. URL http://arxiv.org/abs/1802.05957. 3.2.3, 4, 4.2

[25] Nyukat. nyukat/breast_cancer_classifier. URL https://github.com/nyukat/breast_cancer_classifier. (document), 2.1.1, 2.1, 2.1.2, 2.2

[26] The pandas development team. pandas-dev/pandas: Pandas, February 2020. URL https://doi.org/10.5281/zenodo.3509134. 3.1.3

[27] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 4.1

[28] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2015. URL http://arxiv.org/abs/1511.06434. cite arxiv:1511.06434Comment: Under review as a conference paper at ICLR 2016. (document), 1

[29] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2016. doi: 10.18653/v1/n16-3020. 1

[30] Timo Schick and Hinrich Schütze. BERTRAM: improved word embeddings have big impact on contextualized model performance. *CoRR*, abs/1910.07181, 2019. URL http://arxiv.org/abs/1910.07181. 3.2.1

[31] Max Schubach, Matteo Re, Peter Robinson, and Giorgio Valentini. Imbalance-aware machine learning for predicting rare and common disease-associated non-coding variants. *Scientific Reports*, 7, 06 2017. doi: 10.1038/s41598-017-03011-5. (document), 3.11

[32] Li Shen, Laurie R. Margolies, Joseph H. Rothstein, Eugene Fluder, Russell Mcbride, and Weiva Sieh. Deep learning to improve breast cancer detection on screening mammography. *Scientific Reports*, 9(1), 2019. doi: 10.1038/s41598-019-48995-4. 2.1.2

[33] Yiqiu Shen, Nan Wu, Jason Phang, Jungkyu Park, Kangning Liu, Sudarshini Tyagi, Laura Heacock, S. Gene Kim, Linda Moy, Kyunghyun Cho, and Krzysztof J. Geras. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Medical Image Analysis*, 68:101908, 2021. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2020.101908. URL https://www.sciencedirect.com/science/article/pii/S1361841520302723. 2.1.1

[34] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 2019. doi: 10.1186/s40537-019-0197-0. (document), 1.1, 1, 2.2.3

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. (document), 2.1.1, 3.2.4

[36] Zakia Sultana Ritu, Nafisa Nowshin, Md Mahadi Hasan Nahid, and Sabir Ismail. Performance analysis of different word embedding models on bangla language. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–5, 2018. doi: 10.1109/ICBSLP.2018.8554681. 3.2.1

[37] P Umesh. Image processing in python. *CSI Communications*, 23, 2012. 3.1.3

[38] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL `http://www.jmlr.org/papers/v9/vandermaaten08a.html`. 4.1

[39] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014. 3.1.3, 4.1

[40] Nan Wu, Krzysztof J. Geras, Yiqiu Shen, Jingyi Su, S. Gene Kim, Eric Kim, Stacey Wolfson, Linda Moy, and Kyunghyun Cho. Breast density classification with deep convolutional neural networks. *CoRR*, abs/1711.03674, 2017. URL `http://arxiv.org/abs/1711.03674`. 2.1.1

[41] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, S. Gene Kim, Laura Heacock, Linda Moy, Kyunghyun Cho, and Krzysztof J. Geras. The nyu breast cancer screening dataset v1.0, 2019. URL `https://cs.nyu.edu/~kgeras/reports/datav1.0.pdf`. (document), 2.1.1, 2.1, 2.1.2, 2.2, 3.2

[42] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrzebski, Thibault Févry, Joe Katsnelson, Eric Kim, Stacey Wolfson, Ujas Parikh, Sushma Gaddam, Leng Leng Young Lin, Kara Ho, Joshua D. Weinstein, Beatriu Reig, Yiming Gao, Hildegard Toth, Kristine Pysarenko, Alana Lewin, Jiyon Lee, Krystal Airola, Eralda Mema, Stephanie Chung, Esther Hwang, Naziya Samreen, S. Gene Kim, Laura Heacock, Linda Moy, Kyunghyun Cho, and Krzysztof J. Geras. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Transactions on Medical Imaging*, 39(4):1184–1194, 2020. doi: 10.1109/TMI.2019.2945514. (document), 2.1.1, 2.1, 2.1.2, 2.2