MASTER THESIS and INTERNSHIP REPORT

Finding Balance:
A Study of Class Imbalance in Deep Learning Classification of Breast Cancer

Author:
Ricky WALSH

Company Supervisor:
Mickael TARDY, PhD

Examining Committee:
Simon MALINOWSKI, *Associate Professor*
Alvaro PINA STRANGER, *Associate Professor*

A thesis submitted in fulfillment of the requirements for the degree of Master in Data Science

in the

Department of ISTIC

Université de Rennes 1

France

August 21, 2022

Abstract

Breast cancer is the leading cause of cancer death among women worldwide. The breast screening programmes now active in many countries help to reduce the mortality rate, but result in a large number of images for radiologists to process. Tools based on deep learning have been developed to aid radiologists to diagnose breast cancer from mammograms more efficiently and with better accuracy. However, many of the datasets available to train these deep learning models have a class imbalance problem, i.e., there are fewer images of breasts with malignant lesions than images with no malignancies, which can bias trained models towards the non-malignant class. This report is based on a six month internship at Hera-MI, Nantes, France. During the internship, a systematic study of common techniques for dealing with class imbalanced was carried out on several public and private datasets. Inserting synthetic lesions was also examined as a method to tackle class imbalance. It was found that although class imbalance does indeed shift predictions towards the majority class, models were still able to separate benign from malignant as much as when common class imbalance techniques were applied. Furthermore, synthetic lesions showed significant promise, with improvements in AUC-ROC of 0.02 on an in-distribution test set and up to 0.07 on out-of-distribution test sets.

Résumé

Le cancer du sein est la principale cause de décès par cancer chez les femmes. Le dépistage du cancer du sein peut réduire le taux de mortalité, mais il crée de nombreux images qui nécessitent l'attention des radiologues. Des outils utilisant la technologie d'apprentissage profond sont disponibles depuis récemment pour aider les radiologues à traiter toutes ces images. Cependant, parmi les données utilisées pour créer ces outils, il y a plus d'exemples de seins normaux que de seins présentant des anomalies malignes, ce qui peut causer un biais vers la classe normale dans les modèles prédictifs. Ce rapport présente une étude sur ce problème de déséquilibre des classes, qui a été réalisée lors d'un stage chez Hera-MI, une société d'imagérie médicale à Nantes. Les résultats montrent que bien qu'il y ait du biais en cas de déséquilibre, le modèle peut encore séparer les deux classes aussi bien que lorsque des techniques sont appliqués pour contrer le déséquilibre. Par ailleurs, l'étude démontre l'utilité d'une méthode d'insertion de lésions malignes synthetiques pour équilibrer les classes.

Acknowledgements

I would like to thank everyone at Hera-MI for welcoming me there as part of the team. I was slightly nervous before starting, especially about speaking French with everyone, but several members of the team went out of their way to ensure that I felt involved, even if I did not always understand everything that was happening. Thank you to the other members of the research team in particular, from whose knowledge and experience I learned a great deal. A special thanks goes to my supervisor, Mickael Tardy, who was generous with his time and his help throughout the internship. Whether to give advice, troubleshoot something together, or have an interesting discussion about breast cancer or deep learning, Mickael was always happy to share his considerable expertise. I wish everyone the very best for the future, and look forward to seeing Hera-MI grow into a major provider of AI solutions for medical imaging in France and abroad.

Contents

List of Figures

List of Tables

List of Abbreviations

ACR	**A**merican **C**ollege of **R**adiology
AI	**A**rtificial **I**ntelligence
AUC	**A**rea **U**nder the **C**urve
BI-RADS	**B**reast **I**maging **R**eporting **A**nd **D**ata **S**ystem
CADx	**C**omputer **A**ided **D**iagnosis
CC	**C**ranio**C**audal
CMMD	**C**hinese **M**ammography **D**atabase
DL	**D**eep **L**earning
FDA	**F**ood and **D**rug **A**dministration
FFDM	**F**ull **F**ield **D**igital **M**ammography
FN	**F**alse **N**egative
FP	**F**alse **P**ositive
GAN	**G**enerative **A**dversarial **N**etwork
GPU	**G**raphical **P**rocessing **U**nit
HMI	**H**era-**MI**
IDE	**I**ntegrated **D**evelopment **E**nvironment
MCC	**M**atthews **C**orrelation **C**oefficient
MRI	**M**agnetic **R**esonance **I**maging
MLO	**M**edio-**L**ateral **O**blique
MLP	**M**ulti-**L**ayer **P**erceptron
PR	**P**recision **R**ecall (curve)
ROC	**R**eceiver **O**perating **C**haracteristic
RQ	**R**esearch **Q**uestion
Se	**Se**nsitivity
SMOTE	**S**ynthetic **M**inority **O**ver-sampling **Te**chnique
Sp	**Sp**ecificity
SVM	**S**upport **V**ector **M**achine
TP	**T**rue **P**ositive
TN	**T**rue **N**egative

Preface

When I decided in 2020 to return to university to complete a master's degree, I did not know quite what to expect. Having already worked four years in the data domain, I found myself wanting more. I wanted a challenge. I wanted to explore. I wanted to learn. I pursued the degree primarily to deepen my technical skills and broaden my horizons, believing that this would give me the opportunity to work on more diverse, more complex problems, and hopefully experience the joy of solving them. The innovation and entrepreneurship aspects of the European Institute of Innovation & Technology (EIT) master's programme also appealed to me, perhaps allowing me to take those technical skills and find ways of applying them to solve real-world problems.

I have learned a lot over the past two years, well beyond the subjects taught in the curriculum. The interminable lock-downs during the early days of the Covid-19 pandemic provided me with time for reflection and introspection. New countries and cultures, and diverse groups of international friends, opened my eyes to interesting differences but also significant similarities between us all. Most recently, a year in France, and six months spent at Hera-MI, have satisfied a long-standing wish of mine to immerse myself in the French culture and language. And most importantly, I became a connoisseur of Breton galettes!

During the master's programme I discovered the transformative effect that data and machine learning projects can have on healthcare. This appealed to me tremendously, showing me a way to apply my skills to complex problems to find solutions that might improve the lives of others in some way. Thus, I worked on as many projects in this area as possible as part of my studies, and was lucky enough that one of these projects led to a publication in a scientific journal. This initial experience of research gave me a taste for more, motivating me to seek similar experiences in the internship on which this report is based, and ultimately led to me applying for a PhD that I will begin later this year, on automated detection and segmentation of multiple sclerosis lesions in spinal cord Magnetic Resonance Imaging (MRI).

When I discovered Hera-MI and the work they do in helping to detect breast cancer using artificial intelligence, it was the perfect fit for me. Breast cancer affects so many people and almost everyone knows someone whose life has been changed, or taken, by this disease. I was excited to be part of an effort to catch breast cancer early, and thus to improve patient outcomes. My personal goals for the internship were to gain practical experience in the AI for medical imaging domain, improve my research skills, gain an insight into life in a startup in this area, and, in effect, to do something useful. My main focus during the internship was on investigating techniques to deal with an imbalance between the number of cancerous (malignant) and non-cancerous (benign or normal) cases in datasets used to create models which classify images as to whether the breast shows signs of cancer or not. We plan to submit our research on this topic for publication during the coming months.

This report is divided into four main chapters. Chapter 1 provides a brief background on Hera-MI and the industry context. Chapter 2 discusses the internship as a whole, the goals of the internship, how it progressed, and the format it took. Chapter 3 presents the scientific research undertaken, including a literature review of the state of the art, and the various experiments performed. This makes up the majority of this report, and is based on the text of the journal article to be published. Finally, Chapter 4 offers some reflections on the internship, both from a personal perspective and examining the impact of the six months.

Chapter 1

Company and Industry Context

1.1 AI for Healthcare

Techniques based on Artificial Intelligence (AI) have shown great promise over the last decade. The capabilities of Deep Learning (DL) methods in particular have been underpinned by improvements in hardware and processing power, enabling ever more powerful models to automate complex tasks. One domain in which AI has proved useful is healthcare, in areas as wide as genomics, drug discovery, interpreting electrocardiograms, and, most commonly, medical imaging (Rajpurkar et al., 2022). The use of AI in medical imaging encapsulates several different tasks and many types of medical images. Applications of the technology include diagnosing skin cancer using natural images of skin lesions (Dildar et al., 2021), identifying signs of multiple sclerosis in Magnetic Resonance Imaging (MRI) images of the brain (Shoeibi et al., 2021), or localising bone fractures in an X-ray (Cheng et al., 2021).

There has been much study of these new solutions in the academic literature, and an industry is now growing to commercialise products built on this technology, with global market size projected to grow 300% to $3.2 billion in 2027 by one estimate.[1] This industry is not confined to existing medical technology companies, as tech giants like Microsoft, IBM and Google have also invested significantly in various ventures (Lundervold and Lundervold, 2019), and many startups have been created with a focus on AI for medical imaging (Zhou et al., 2021). According to a database of the American College of Radiology's (ACR) Data Science Institute[2], nearly 200 products based on AI for medical imaging have been approved by the FDA (Food and Drug Administration) in the USA. More than 10% of these licensed products are focused on breast imaging, which gives an indication of the significance of breast cancer, and the level of innovation dedicated to tackling this disease.

1.2 Breast Cancer

Breast cancer is the leading cause of cancer death among women worldwide, and was responsible for an estimated 680,000 deaths in 2020 (Sung et al., 2021). As a result, many countries have introduced breast screening programs which, along with early treatment, can reduce mortality rates by 26-38% (Broeders et al., 2012; Mandelblatt et al., 2016). Mammography is the most widely used imaging modality in screening programs (Peintinger, 2019).

A mammogram is an X-ray image of the breast. The breast is compressed and spread out, then beams of low-dose radiation are passed through and captured in

[1] https://www.researchandmarkets.com/reports/5337141/artificial-intelligence-in-medical-imaging-market - accessed 15 August 2022

[2] https://aicentral.acrdsi.org/ - accessed 7 August 2022

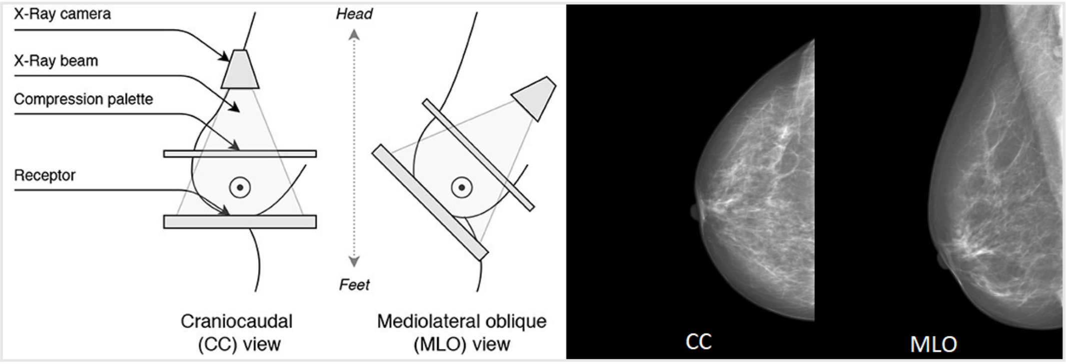

FIGURE 1.1: Illustration of the two most common views acquired with mammography. Left: diagram explaining acquisition (Tardy and Mateus, 2022). Right: Example of two views of the same breast.

a receptor at the other side. Areas of higher density, such as a mass or a build-up of calcium deposits, absorb more of the passing x-rays and appear brighter in the resulting image, allowing radiologists to identify potential malignancies. Typically, two images of each breast are acquired from different angles, Craniocaudal (CC) and Mediolateral Oblique (MLO), illustrations of which are given in Figure 1.1. Further imaging, such as ultrasound, MRI, tomosynthesis or further mammograms, can be used when a lesion is suspected, if the results are unclear from the first set of mammograms, or for women with a higher risk of developing breast cancer (Rebolj et al., 2018; Kriege et al., 2004; Bevers et al., 2018).

Breast cancer screening programs produce a large number of images requiring the attention of radiologists for diagnosis, which can be a tedious, time-consuming and costly process (Geller et al., 2009; Ribli et al., 2018). There are already signs of professional burnout among breast imaging radiologists (Parikh, Sun, and Mainiero, 2020), which could be compounded by the large number of images produced as well as expected increases in workforce shortages over the next number of years (Wing and Langelier, 2009). Moreover, increasing an already high workload for radiologists may lead to an increase in errors (Brady, 2017).

Computer-aided Diagnosis (CADx) tools were proposed to deal with these problems and the use of these increased in the early 2000s, albeit with mixed results (Doi, 2007; Lehman et al., 2015). More recently, automated diagnosis systems have been developed based on DL models which learn to distinguish between benign and malignant lesions in breasts from many retrospective examples. These models have shown improved performance and large clinical studies have demonstrated their usefulness as an aid to radiologists in screening mammography (McKinney et al., 2020; Schaffter et al., 2020; Conant et al., 2019; Rodriguez-Ruiz et al., 2019).

1.3 Company Overview

The internship took place in Hera-MI, a medical imaging company based in Nantes, France. The company was set up in 2017 with the aim of improving the early detection of breast cancer by creating AI solutions for mammography. Their primary product, Breast-SlimView, is an aid to radiologists in mammography screening. Through a patented process the company calls "negativation", normal areas of the breast are masked automatically (see Figure 1.2), thus focusing the attention of

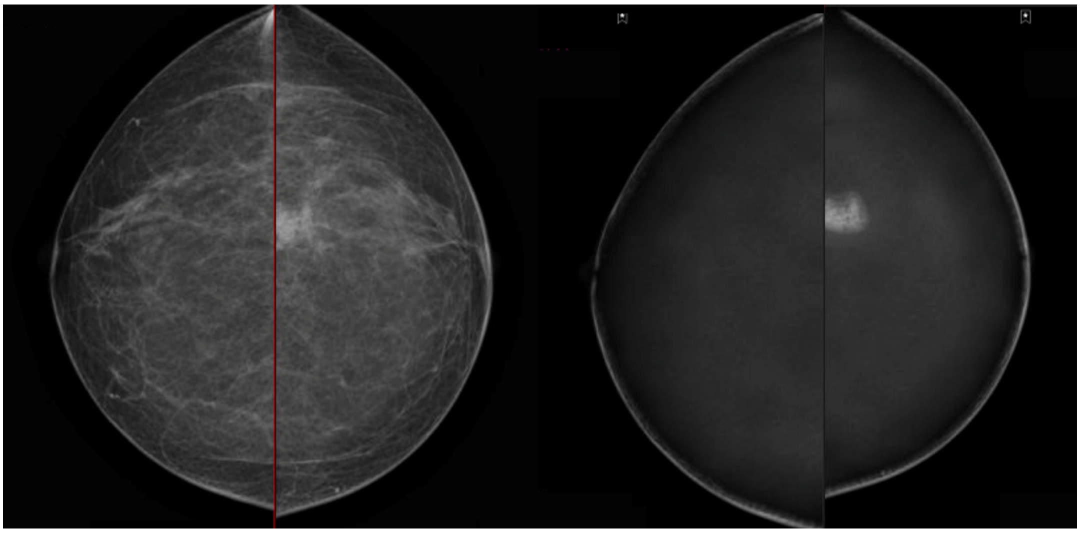

FIGURE 1.2: Demonstration of Breast SlimView software. Left: Original CC views for two breasts of one patient. Right: Output of Breast SlimView - normal areas are masked, and malignant mass is highlighted.

radiologists on abnormalities. The tool helps to reduce the amount of time needed by clinicians in screening mammography, while also improving detection rates.

The company follows a typical business model for this domain in selling the software to radiologists and clinics, both directly and through distributors. Two key partnerships with Fujifilm and Medecom, who incorporate the software into the diagnostic workstations they sell, enable the company to reach more radiologists, and thus benefit more patients. Competitors propose other CAD tools in mammography based on AI, both in France (Therapixel) and abroad (e.g., Screenpoint, Kheiron, and i-CAD), but Hera-MI point to their unique value proposition in negativation, rather than inserting bounding boxes around lesions, which allows radiologists to avoid an information overload.

Early funding from angel investors, regional initiatives, and innovation prizes allowed the company to develop in the early years, enabling the significant research required to create diagnostic software in the medical domain. In 2019, the company then became the first French company to receive CE certification for an AI solution to diagnose breast cancer in mammography. Now, after several years of success with Breast Slim-View and a solid base in France, Hera-MI is expanding its network across Europe, and is in the process of seeking FDA approval. Moreover, the company has broadened its research focus to the detection of different cancers, e.g., prostate cancer, and to different imaging modalities, with the tool already capable of processing Digital Breast Tomosynthesis (DBT) and breast MRI being an active area of development. Indeed as the AI for medical imaging industry matures, companies will compete increasingly on the breadth of their offering to radiologists.

Chapter 2

Internship Journey

2.1 Internship Context

For the duration of the internship, I was a member of the research team in Hera-MI. The company supervisor of the internship, Mickael Tardy, is Chief Scientific Officer and leads the research team. The team has eight members and so accounts for a significant portion of the company's overall headcount of 20. The other members of the research team work on related topics in AI for classification and segmentation of medical images, as well as general image processing, and their experience and helpfulness provided me with valuable insights throughout the internship.

The research team has a weekly meeting where each member presents some of the work they did that week, which I found to be useful during the internship, both to learn from the other researchers, and to share my progress and get feedback and ideas. Being in the office several days each week next to my supervisor and the other researchers also allowed me to quickly share any thoughts or questions I had. Moreover, my supervisor and I met several times throughout the internship when I wanted to share in-depth results or ask more detailed questions. Most researchers at the company work in a hybrid fashion, working remotely for part of the week and spending the other days at the office. I found this setup suited me well, as working remotely allowed more focus for bibliographic research or complex coding tasks, whereas being at the office provided the opportunity to easily ask colleagues for advice and helped me to feel part of the team.

2.2 Internship Mission

The internships that take place at Hera-MI vary in scope and in nature, depending on the needs of the company at the time, as well as the profile and interests of the intern. Some, including mine, take a relatively narrow focus on a single research topic, while others work on research questions that arise from the broader team during the course of the internship. After starting the internship, I began with exploratory analysis on breast imaging datasets while my supervisor and I discussed potential research topics, until we settled on the problem of class imbalance.

Frequently, the public and private datasets that the company works with have an over-representation of images without signs of cancer, relative to images with cancerous lesions. In a classification problem, this is referred to as class imbalance. Although this may not be a problem if there were millions of malignant images available for use, this is not the case as, in fact, most datasets in the area comprise thousands of images. Therefore, researchers must attempt to learn generalisable characteristics of malignancy from a small number of cancerous cases. This is compounded by somewhat noisy labels, for example, a malignant lesion in a breast might show up in the MLO view and not the corresponding CC image. The image labels are dependent on the breast, though, so this latter CC image is likely to be included as malignant during training without any visible signs of malignancy.

The internship mission was thus to investigate how best to utilise these relatively small sets of images under situations of class imbalance. Upon reviewing the literature (detailed in Section 3.2), it was evident that there were many techniques available, but no systematic review of their effects on a breast cancer classification problem. Furthermore, other studies of class imbalance showed that the effect of techniques can vary from domain to domain. Therefore, the agreed internship mission was to conduct a set of systematic experiments on several heterogeneous datasets using the most commonly used class imbalance techniques. We further planned to explore the effect of inserting synthetic malignant lesions into images of normal breasts as an alternative technique to balance the malignant and benign classes. This innovative synthetic lesion methodology had been developed in-house as a way to train segmentation models, and the team was keen to explore further uses.

2.3 Planning and Initial Steps

Figure 2.1 presents an overview of the progression of the internship. Exploratory analysis and a literature review were essential first steps in getting familiar with the domain, the state of the art, and the available data. I was able to apply visualisation techniques I had learned during my work as a data analyst, as well as image processing and programming skills that I developed during the master's degree, while at the same time deepening my knowledge of the domain with, for example, a one week company training on mammography and breast MRI.

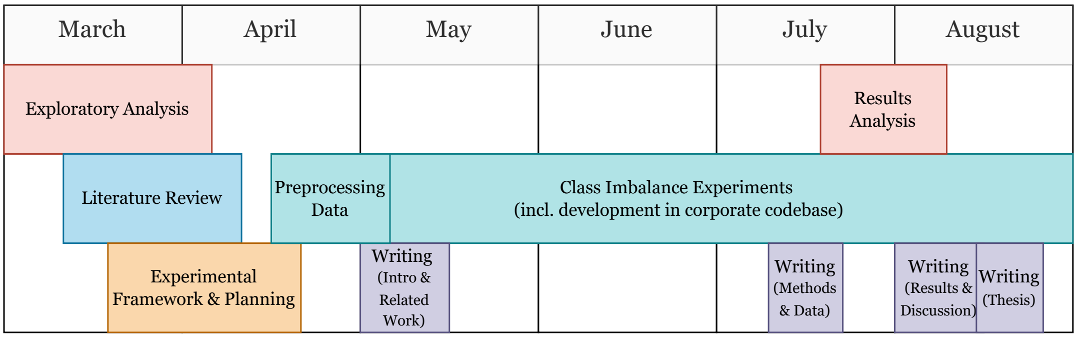
March
April
May
June
July
August
Exploratory Analysis
Results Analysis
Literature Review
Preprocessing Data
Class Imbalance Experiments (incl. development in corporate codebase)
Experimental Framework & Planning
Writing (Intro & Related Work)
Writing (Methods & Data)
Writing (Results & Discussion)
Writing (Thesis)

FIGURE 2.1: High-level overview of internship timeline.

Figure 2.2 is an example of the output of my work at that time. The INBreast and CMMD (Chinese Mammography Database) datasets in these plots will be introduced in more detail in the body of the report (see Section 3.4). We can see that, initially, the three datasets have quite different distributions over the range of intensities. This is important because a machine learning model trained on data from one distribution may fail to generalise to different distributions. We see in the second plot that histogram normalisation can partly resolve this, reducing the differences between datasets. Histogram normalisation involves stretching the histogram of intensities in each image to ensure the full range of available intensities is used. A further insight that we gained from this analysis is that the CMMD images are already quite stretched across the intensity range, and so histogram normalisation has less of an effect on this distribution than for other datasets. This can impact the choice of normalisation applied before training a new model, and indeed we later confirmed

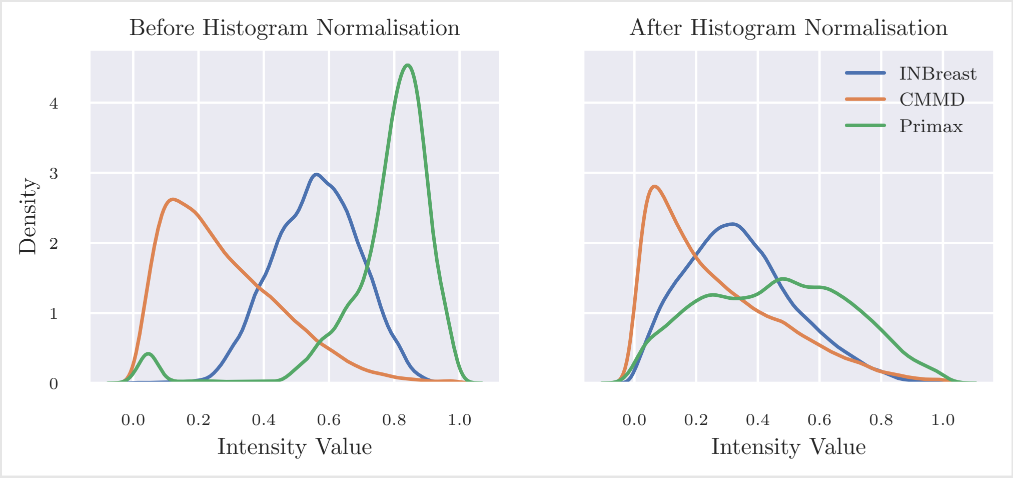

FIGURE 2.2: Comparison of intensity distributions of three datasets, and the effect of histogram normalisation. Shown are the average intensity distributions across all images in each dataset.

that this normalisation did not greatly affect classification results for CMMD data. In general, insights gathered during this initial exploration period can benefit the team by informing decisions of the other researchers when using those datasets, and thus reducing necessary research times.

As mentioned in Section 2.2, my supervisor and I agreed that the experimental framework would be to systematically carry out a series of experiments applying several class imbalance techniques, which I identified in the literature, to several different datasets, including most of the publicly available Full-Field Digital Mammography (FFDM) datasets. We initially identified a larger set of techniques, and I started with the most commonly applied, knowing that we may not be able to complete everything in a matter of months. The main constraint on the number of possible experiments was the significant time required to train the models, as each experiment might run for one day or for several weeks, depending on the dataset, the method applied, and the hardware used. Therefore, as the experiments progressed, we reviewed the time taken for each experiment and were better able to forecast the time needed for the remaining experiments, to ensure we would have a complete, valid set of results that could be published.

The time spent on class imbalance experiments included 1) implementing each technique, 2) developing evaluation methodologies, 3) training the models, 4) assessing results, and 5) exploring some of the hyperparameters and pre-processing techniques. Although in Section 3.5 I present results of 14 experiments, the number of training runs I launched is closer to 100, with extra runs to explore configurations, or to troubleshoot when certain experiments were not working as well as expected. A Graphical Processing Unit (GPU) is essential for extensive deep learning experiments of this kind to accelerate the training process. Many members of the research team are working on deep learning research, so the GPU resources are shared. For most of the internship, I had access to one GPU and occasionally two, so it was important to plan to keep the available GPU occupied as much as possible, allowing more experiments to be run over the full period. Therefore, when model training for each experiment was finished, the code and configuration for the next experiment was ready to run, and this helped to create mini-deadlines throughout the internship. Then, when time allowed, I progressively wrote the sections of the paper we aim at submitting for publication.

Chapter 3

Study on Class Imbalance

3.1 Introduction

As stated in Section 1.2, automated diagnosis of breast cancer based on DL models is becoming more widely used and showing promising results. However, datasets used to train these models are often highly imbalanced, i.e., they contain more benign samples than malignant, as most women who undergo mammography screening do not have breast cancer. This problem of different proportions of each class of interest is called class imbalance, and numerous studies have shown that it can be detrimental to the performance of a classification model (Japkowicz and Stephen, 2002; Buda, Maki, and Mazurowski, 2018). The imbalance can cause models to be biased towards the majority class, which is of particular concern in mammography screening as this may lead to models being more likely to predict images as being benign, potentially leading to missed cancers.

Many techniques have been proposed to tackle the effect of class imbalance, but their impacts can vary depending on the complexity of the task and the distribution of the dataset (Buda, Maki, and Mazurowski, 2018; Johnson and Khoshgoftaar, 2019). In other words, a method that is well suited to optical character recognition, for example, may be unsuitable for cancer classification. This motivates the need for a broad study of these techniques applied to breast cancer diagnosis using various mammography datasets, to determine whether certain techniques are best suited to this task.

The task considered is to classify whether a breast contains cancerous lesions or not based on the information presented in a mammogram. Although the focus of this study is on mammography, many other medical imaging tasks share the same main characteristics, namely, high-resolution images, a clinical context where the trade-off between sensitivity and specificity is important, and often a high level of imbalance between classes. Thus, the results of this study could have wider applicability for medical imaging in general.

The aim of this study can be summarised in the following Research Questions (RQ):

RQ1 What effect does class imbalance in various mammography datasets have on cancer classification performance and the generalisability of a deep learning model?

RQ2 How do common techniques for tackling class imbalance compare for cancer classification in mammography?

RQ3 To what extent does the addition of images with synthetic lesions during training improve classification performance?

To answer these questions, extensive experiments were conducted on several mammography datasets with different data distributions and levels of class imbalance, including two recently released datasets, Chinese Mammography Database (CMMD) (Cui et al., 2021) and VinDr-Mammo (Nguyen et al., 2022), as well as the popular INBreast dataset (Moreira et al., 2012). A selection of the most popular techniques for handling class imbalance were chosen from the literature and applied to evaluate their effects on classification performance and generalisability of a standard Convolutional Neural Network (CNN) used for breast cancer classification.

This chapter is structured as follows. Section 3.2 discusses related work in automated cancer diagnosis and studies on class imbalance for DL. Section 3.3 details the models, class imbalance techniques, and experiments carried out in this study, while Section 3.4 presents the datasets used. The results of the experiments are presented in Section 3.5. Finally, Section 3.6 contextualises and summarises the key findings from the experiments and their implications.

3.2 Related Work

3.2.1 Breast Cancer Classification

The current state of the art in automated breast cancer classification from mammograms involves CNNs trained to classify full resolution images (Wu et al., 2019), subsets of images in the form of patches (Choukroun et al., 2017), or a combination of the two (Shen et al., 2019). Areas showing promising results in the domain include Multi-Task Learning (Tardy and Mateus, 2022), where supplementary information in the data annotations such as breast density and the Breast Imaging-Reporting and Data System (BI-RADS) risk rating can be used during training to help the model to converge. Multiple instance learning has also proved useful as it can provide a way of localising abnormalities in images using only image-wise malignancy labels rather than pixel-wise ground truth annotations of abnormalities (Choukroun et al., 2017; Bakalo, Ben-Ari, and Goldberger, 2019).

The majority of the studies in this area, however, use datasets with a higher number of benign samples than malignant. Although some common measures are used to counteract this imbalance, such as over-sampling and class weighting, the question remains as to how impactful class imbalance is in this domain, and how well different techniques resolve the problem.

3.2.2 Effects of Class Imbalance

Study of the class imbalance problem is not new in the literature. As early as 1993, Anand et al. (1993) showed that when training a shallow neural network with back-propagation on an imbalanced dataset, the gradient contribution of the majority class dominates that of the minority class, which leads to slow convergence of the error for the minority class. As machine learning techniques became more widely studied, a significant amount of research was dedicated to class imbalance, with a number of workshops and special issues in the early 2000s (Chawla, Japkowicz, and Kotcz, 2004). More recently, Li, Kamnitsas, and Glocker (2021) used two different CNNs for the segmentation of brain tumours and other anatomical structures in various datasets, testing different levels of imbalance. The authors demonstrated that for a higher level of imbalance, the model overfits more on the majority class.

Japkowicz and Stephen (2002) demonstrated that class imbalance can be more detrimental for more complex tasks. The authors also found, however, that imbalance was less problematic for larger datasets, and that the impact varied depending on the classification algorithm used. Mazurowski et al. (2008) focused on the task of medical diagnosis, training a Multi-Layer Perceptron (MLP) to classify breast cancer from manually extracted features of masses in mammograms. The authors found that increasing class imbalance in the training set was generally associated with a lower test performance, and that the impact was stronger for a real breast cancer dataset than for an artificial dataset, likely due to increased complexity of the data.

Buda, Maki, and Mazurowski (2018) studied class imbalance in the context of CNN models trained to recognise digits using the MNIST dataset, or to classify ten everyday objects from the popular CIFAR-10 image dataset. The authors confirmed that class imbalance remains a problem even for advanced models such as CNNs, leading to reduced classification performance. Moreover, class imbalance was more impactful on the more complex CIFAR-10 dataset than for MNIST, providing further evidence that the effect depends on the complexity of the dataset.

In summary, while the class imbalance problem has been well studied over the past twenty years and its impact has been observed for many domains and datasets, the effect of class imbalance can vary considerably depending on the complexity of the task at hand and the characteristics of the dataset. Moreover, a broad study on class imbalance in automated cancer detection in mammography across several datasets has not been carried out. Although Bria, Marrocco, and Tortorella (2020) explored similar questions, they used a dataset of 14 × 14 pixel patches extracted from the INBreast dataset with the aim of classifying micro-calcifications, which reveals little about the effect when processing whole images, as the task becomes more complex.

3.2.3 Methods for Dealing with Class Imbalance

Common Techniques

Techniques proposed to deal with the class imbalance problem are often categorised as algorithm-level, data-level, and a combination of the two (Chawla, Japkowicz, and Kotcz, 2004; Krawczyk, 2016). The most widely used algorithm-based techniques involve changes to the loss function, for example, weighting by the inverse of the class proportions such that the contribution from the majority class and minority class are balanced (Johnson and Khoshgoftaar, 2019). This has previously been used for studies in deep learning in mammography (Shen et al., 2019; Zhu et al., 2017), though Bria, Marrocco, and Tortorella (2020) found it less effective than over-sampling in their study.

The primary techniques in the data-level group include sampling, where either the minority class is over-sampled or the majority class is under-sampled, creating an artificially balanced dataset. There are some potential drawbacks, in that under-sampling may remove important, informative examples, and over-sampling may lead to overfitting (Chawla, Japkowicz, and Kotcz, 2004). However, in the context of CNNs, the extensive use of data augmentation during training often reduces the risk of overfitting in general (Taylor and Nitschke, 2019), and so might help to avoid overfitting on over-sampled minority examples. Despite these sampling methods being some of the earliest techniques addressing the class imbalance problem, they remain popular, perhaps due to their simplicity and ostensible effectiveness. For example, one review of the use of CNNs in mammography (Abdelhafiz et al., 2019) discusses the problem of class imbalance and mentions only these two techniques to counter the imbalance effect.

In comparing several different sampling strategies, Bria, Marrocco, and Tortorella (2020) found that over-sampling the malignant class was the most effective for their problem, along with directed under-sampling of the normal samples. This latter technique, also known as hard sample mining, involves selecting only the samples in the training set on which the model is performing poorly. Qu et al. (2020) found over-sampling and under-sampling both to be effective in reducing the class imbalance effect for classification of X-rays, although they evaluated performance only for one threshold on the softmax outputs, rather than assessing the overall model performance with a metric such as the Area Under the Curve of the Receiving Operating Characteristic (AUC-ROC).

The most common methods used to tackle class imbalance in medical imaging studies include the aforementioned over-sampling, under-sampling, and class weighting. However, the effectiveness of these techniques varies depending on the task being performed and the dataset (Johnson and Khoshgoftaar, 2019; Buda, Maki, and Mazurowski, 2018). Therefore, this study examines how effective these common techniques are in dealing with class imbalance in three heterogeneous mammography datasets.

Synthesizing Images

To avoid overfitting on over-sampled minority examples, further data augmentation can be applied (Parmar et al., 2018), and one step further is to create new synthetic images for the minority class. Generative Adversarial Networks (GAN) are often used for this latter task, and several studies used this as an additional type of data augmentation during training, alongside typical image flipping, rotation, etc. They have been shown capable of producing realistic images at a low-resolution, and the added synthetic images helped to improve model performance in studies of classification of liver lesions (Frid-Adar et al., 2018), chest X-ray abnormalities (Ali Madani et al., 2018), and breast masses (Alyafi, Diaz, and Martí, 2020), and in a study on segmentation in brain scans (Bowles et al., 2018). GANs have also shown some success for explicitly tackling class imbalance in mammogram classification (Wu et al., 2018), although the authors did not compare their approach to other common class imbalance techniques.

In all of the above cited studies, the GANs were trained to produce lower resolution images or patches between 64 × 64 and 256 × 256 pixels, and so it is unclear how well a GAN would handle whole high resolution mammograms where lesions might appear in less than 1% of the pixels in the image. Although Korkinof et al. (2018) succeeded in synthesising higher resolution (1280 × 1024) mammograms using a GAN, the model needed 450,000 images for training, and the training process remained relatively unstable, with several training runs failing unexpectedly. Moreover, many studies on GANs dealing with breast lesions require explicit pixel-wise ground truth for lesion locations, that are rarely available in clinical practice.

In this study, an alternative method of artificial lesion generation is used to insert masses, calcifications, and architectural distortions into benign images to tackle class imbalance. The mass generation is based on a computational model by Sisternes et al. (2015), and was previously shown to be useful as a data augmentation technique in training a deep learning model for mass detection (Cha et al., 2019). The method used here was used previously for lesion segmentation (Tardy and Mateus, 2021), to deal with the absence of ground truth lesion locations.

There are three main advantages of this method compared to the GAN techniques discussed earlier. Firstly, they can be used with images of any resolution, in particular high resolution images. Secondly, ground truth lesion locations are not required, which are expensive and rare. Finally, whereas a GAN is confined to the distribution of the dataset at hand, controlling the artificial lesion characteristics with the current method allows domain knowledge on the appearance of different lesions to be incorporated, regardless of whether certain lesion types or shapes are represented in the training dataset. Moreover, this method could be used to balance datasets where there is a known shortage of certain types of masses, for example.

In summary, the contributions of the current study are: an analysis of the effect of class imbalance on breast cancer classification using high resolution images from several different mammography datasets, a systematic comparison of the most popular methods for dealing with this class imbalance problem, and a novel use of synthetically generating abnormalities as an alternative to over-sampling the malignant class.

3.3 Methods

3.3.1 Dealing with Class Imbalance

Several of the most popular methods for addressing class imbalance were examined, namely class weighting, under-sampling, and over-sampling. A synthetic lesion generation technique, referred to here as *Artifacting*, was also explored and employed in a novel way for tackling class imbalance in classification problems.

Class Weighting

One of the simplest approaches is to apply a higher weight to the minority class during training when calculating the loss. Eq. 3.1 below shows how the weights were calculated. These were global weights, i.e., using the total numbers of minority and majority samples in the training dataset rather than, for example, calculating the ratio for each mini-batch during training. This decision was based on the small batch size of 8 images relative to the large class imbalance ratio of 19:1 (i.e., 19 benign images for every one malignant image) in the VinDR dataset. Thus, re-calculating class weights dynamically for each batch would mean the largest weight any minority sample could possibly receive would be 7, which would not balance the loss contributions of both classes.

$$w_{minority} = \frac{\#\, of\ Majority\ Samples}{\#\, of\ Minority\ Samples}, \quad w_{majority} = 1 \tag{3.1}$$

Under-sampling

With this approach, a fixed random sample of the majority class is taken before training so that the number of minority and majority examples are balanced. For datasets with large imbalance between classes, this leads to a dramatic reduction of the training data size, e.g., a 90% reduction of the training data size for the VinDR dataset, having a benign to malignant ratio of 19:1. For cases where the class imbalance is less severe, like the other two datasets used, under-sampling removes fewer samples and thus less potentially valuable information is lost.

Over-sampling

For this method, each majority class example is seen exactly once in every epoch whereas minority class examples can be seen multiple times. More specifically, each batch is first half-filled with unseen majority examples and then minority examples are randomly selected to complete the rest of the batch. The epoch is complete when all majority examples have been seen.

Artifacting

The final method involves inserting synthetic malignant lesions, or artifacts, into benign images during training to balance the benign and malignant classes. In the current study, this method was used only for datasets where there are more benign samples than malignant (i.e., VinDR and HMI), as if a malignant lesion is inserted into a benign image it can be considered malignant, but inserting a benign lesion into a malignant image does not change the class of the image, as the source of malignancy would still be there.

Three types of malignant lesions were inserted, namely masses, calcifications, and architectural distortions, examples of which are shown in Figure 3.1 below. The methodology underlying the generation of these synthetic lesions has been described in detail by Tardy (2021) and has previously been employed for lesion segmentation (Tardy and Mateus, 2021). A brief description is given in the following paragraphs.

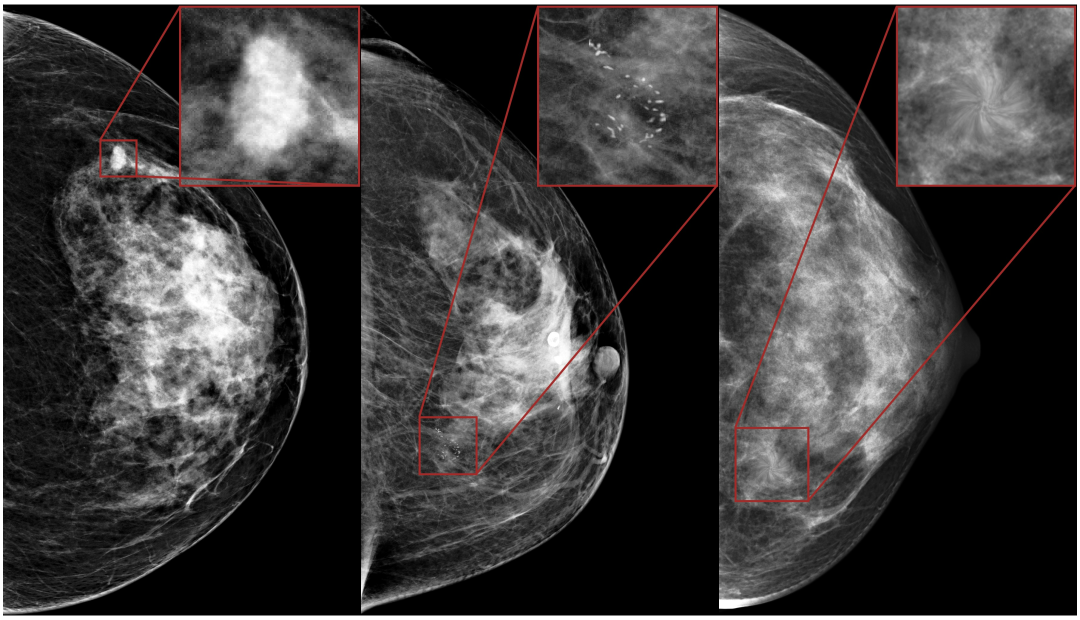

FIGURE 3.1: Examples of images from HMI dataset with synthetic lesions inserted (from left to right: mass, cluster of calcifications, architectural distortion).

The mass generation is based on an independent implementation[1] of the computational model designed by Sisternes et al. (2015). This method uses a stochastic Gaussian random sphere model to generate synthetic masses and then adds spicules to these with an iterative branching algorithm. The authors showed that both radiologists and CAD tools had difficulty in separating real masses from the generated synthetic masses.

[1] https://breastmass.readthedocs.io/en/latest/

A calcification is an accumulation of hardened calcium in the breast tissue. While most calcifications seen in mammograms are benign, some types of calcification are more indicative of malignancy such as smaller calcifications and clusters (Muttarak, Kongmebhol, and Sukhamwang, 2009). Therefore, malignant clusters were imitated by inserting localised regions of small bright spots where each calcification has a diameter of between 0.25mm-1mm, formed in round or elliptical groups of high intensity pixels.

An architectural distortion, defined as a distortion of the breast parenchymal architecture without a definable mass (Bahl et al., 2015), is another type of finding which can be indicative of malignancy. The synthetic distortions were designed to simulate the case where these manifest in a mammogram as a twisting or compression of the tissue in a localised region. These artifacts were thus created with a local non-linear geometric transformation from scikit-image[2].

These three types of synthetic lesion were inserted in a randomised fashion during training. Firstly, each training mini-batch of eight images was half-filled with benign samples. Then, two real malignant samples were added, and the remaining two places were filled by randomly selecting two of the benign samples in the batch and inserting one, two, or three random synthetic lesions. This means that for VinDR, which has a benign to malignant ratio of 19:1, there is both over-sampling of the real malignant examples in addition to the synthetic malignant examples. Ensuring that there are both real and synthetic examples in each batch was found experimentally to yield the best results.

3.3.2 Experimental Framework

The experimental methodology involved applying all of the aforementioned techniques for tackling class imbalance to each of the datasets and training a deep learning model for classification.

Pre-processing

The images were pre-processed before being used for training or testing a classification model. Background noise and labels were first removed from the images by locating the breast as the largest object on the image and setting the other pixels to zero. Images of right breasts were flipped so that all breasts were located at the left side of the images. The height was cropped to the size of the breast, and re-sized to 2,048 pixels. The aspect ratio was maintained between the height and width, but the right side of the image was padded with zero intensity pixels to give a square 2048×2048 image. Finally, the images were scaled to values between 0 and 1, and histogram normalisation was used to enhance the contrast of the images.

Training

A ResNet-22 architecture, which has previously been used for breast cancer classification (Wu et al., 2019) was used as the deep learning classifier. The same version was used in this study as in the encoder block of Tardy and Mateus (2022), with five residual blocks and an increasing number of filters (16, 32, 64, 128, 256), and using separable convolutions and instance normalisation as in Tardy and Mateus (2022). During training, the images were augmented in a randomised fashion, including

[2] https://scikit-image.org/docs/stable/auto_examples/transform/plot_swirl.html

vertical flipping and translation (< 100 pixels), rotation (± 10 degrees), and zooming, as well as inpainting random patches in the image similar to Zhou et al. (2019).

For each experiment, the model was trained from scratch for 100 epochs with a categorical cross-entropy loss, using the Adam optimiser with a learning rate of 5×10^{-4}, informed by previous experiments on the same data and a small number of tests during this study. The batch size was set to 8, which was the maximum possible without exceeding available RAM during training. The training times of the 14 experiments varied between 1-8 days each, and so extensive hyperparameter tuning for each experiment was not feasible. During training, the Area Under the Curve of the Receiver Operating Characteristic (AUC-ROC) was calculated on the validation set, and the five models which achieved the best AUC were selected. Metrics on the test sets were ultimately calculated for each of these five models, and the mean score for each experiment is reported in Section 3.5.

Evaluation

When evaluating predictions on the test set, the breast-wise predictions were assessed, similar to Stadnick et al. (2021). For many of the patients in the datasets, there were two images of each breast, one from each of the CC and MLO views. Signs of malignancy might appear on only one of these views, so the predictions of each view were combined at test-time by taking the average of the model's predicted probability of malignancy for each breast.

The performance of each trained model was evaluated on a test set from the same distribution as the training set, as well as on the other out-of-distribution test sets, exploring the generalisation capabilities of each method. The primary evaluation metric was AUC-ROC. ROC is a plot of sensitivity vs. (1-specificity) for many operating points of the model, i.e., thresholds on the softmax outputs used to decide whether to classify an example as malignant or benign. AUC-ROC captures the information in this curve in one single metric, allowing the overall quality of the model to be evaluated, without choosing a specific threshold. One of the benefits of this for the current study is that setting a different threshold for a trained model is itself a technique for tackling class imbalance (Buda, Maki, and Mazurowski, 2018; Johnson and Khoshgoftaar, 2019), allowing for a different balance between sensitivity and specificity to be achieved. Therefore, assessing the overall model performance is more informative here than performance at one single threshold.

There has been some criticism of AUC-ROC for classification with imbalanced classes, in favour of the alternative Area Under the Precision-Recall Curve (AUC-PR) (Saito and Rehmsmeier, 2015). On the other hand, some studies (Bradley, 1997; Boughorbel, Jarray, and El-Anbari, 2017) have found AUC-ROC to perform well under class imbalance. The main reason, however, that AUC-ROC was chosen for the current study is that the metric is commonly used in studies of breast cancer classification and segmentation, allowing comparisons to other works (Shen et al., 2019; Wu et al., 2019; Tardy and Mateus, 2022).

Model predictions were also assessed with standard metrics at an operating point of 0.5. Sensitivity (Se) and Specificity (Sp) were calculated to evaluate how biased each trained model was towards malignant or benign samples. Matthews Correlation Coefficient (MCC) was used to have a single evaluation metric which incorporates information from True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). Several studies have shown this to be an informative metric in the study of class imbalance problems (Chicco and Jurman, 2020; Boughorbel, Jarray, and El-Anbari, 2017). The formulae for calculating these three metrics are shown below. To get an estimate of the standard error of the metrics for each experiment, bootstrapping (Efron and Tibshirani, 1993) was used with 1,000 samples.

$$Se = \frac{TP}{TP+FN}, \quad Sp = \frac{TN}{TN+FP}, \quad MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}}$$

3.4 Data

Several heterogeneous public datasets and a private dataset were used, each based on different geographical locations and with different levels of class imbalance, to assess the performance of class imbalance techniques. Each of the datasets contains high resolution Full Field Digital Mammography (FFDM) images.

The first dataset used was VinDr-Mammo (Nguyen et al., 2022), consisting of 20,000 images from 5,000 patients. The images were collected in two hospitals in Vietnam, Hanoi Medical University Hospital and Hospital 108, using mammography systems from three vendors (Siemens, Planmed, and Giotto). This dataset presents a large class imbalance, where only 998 (5%) images are labelled as malignant (BI-RADS 4 or 5), and reflects a typical mammography screening scenario where most of the images acquired are normal or benign. The dataset authors have already split the data into training (16,000) and test (4,000) images to ensure consistency in results reported. Hereafter, these will be referred to as $\mathcal{D}_{\mathrm{VinDR}_{\mathrm{train}}}$ and $\mathcal{D}_{\mathrm{VinDR}_{\mathrm{test}}}$, respectively.

The second public dataset used for training was the relatively new Chinese Mammography Database (CMMD) (Cui et al., 2021), which contains 3,728 images of 1,775 patients. The dataset was published by the South China University of Technology, and the images were obtained using a GE Senographe DS mammography system. In this dataset there are more malignant images than benign, reflecting the diagnostic clinical context, where suspicious findings have been identified and further imaging is required to diagnose whether they are malignant or benign. The dataset was split into training (2,982 images – 80%) and test (746 images – 20%) sets, stratifying by class and ensuring that the images for a single patient remained in a single set. These datasets are denoted as $\mathcal{D}_{\mathrm{CMMD}_{\mathrm{train}}}$ and $\mathcal{D}_{\mathrm{CMMD}_{\mathrm{test}}}$.

The third and final dataset used for training was a private dataset of 3,851 images, containing mammograms from four different vendors, namely Fujifilm, GE, Hologic, and Planmed. A modest class imbalance is present in this dataset with benign images accounting for 70% of the training dataset. A test split previously created within the company containing 504 images was used, allowing comparisons to other results internally. These datasets are denoted as $\mathcal{D}_{\mathrm{HMI}_{\mathrm{train}}}$ and $\mathcal{D}_{\mathrm{HMI}_{\mathrm{test}}}$.

For each of the datasets, the training set was further split into separate training and validation sets, using the validation set to tune hyper-parameters and to find the epoch of the "best" models during training. These splits were stratified based on class, and images for each patient appeared in only one of the splits. A summary of the number of images in each split for each dataset is presented in Table 3.1, as well as the class representations.

For each assessed technique for tackling class imbalance, models were separately trained on each of the three above datasets, and the performance was assessed on each of the test sets. Taking a model trained on $\mathcal{D}_{\mathrm{VinDR}_{\mathrm{train}}}$ and testing it on $\mathcal{D}_{\mathrm{CMMD}_{\mathrm{test}}}$, for example, allows the assessment of the generalisation capability on new datasets. Finally, the popular public INBreast (Moreira et al., 2012) dataset was used solely for testing performance across all experiments, allowing the comparison to other

TABLE 3.1: Summary of the number of images in each dataset and subset by class.

Dataset	Training		Validation		Test		Total
	Normal/Benign	Malignant	Normal/Benign	Malignant	Normal/Benign	Malignant	
VinDR	13286 (95%)	704 (5%)	1924 (96%)	86 (4%)	3802 (95%)	198 (5%)	20000
CMMD	678 (30%)	1574 (70%)	216 (29%)	530 (71%)	218 (29%)	528 (71%)	3744
HMI	1952 (70%)	856 (30%)	402 (75%)	133 (25%)	257 (51%)	247 (49%)	3847
INBreast					310 (76%)	100 (24%)	410

studies in the area (Stadnick et al., 2021). This dataset consists of 410 FFDM images taken with a Siemens mammography system.

The task considered by this study is to classify images as malignant or non-malignant (which could include normal cases or benign lesions). In most cases, the BI-RADS rating from the American College of Radiology (ACR) was used to determine malignancy, where the scores are categorised as follows: 1 - negative or normal, 2 - benign, 3 - probably benign, 4 - suspicious for malignancy, 5 - highly suggestive of malignancy, and 6 - known biopsy-proven malignancy. The rating for each image was binarised by considering BI-RADS 1, 2, and 3 to be non-malignant and ratings 4, 5, and 6 to be malignant. In the case of the CMMD dataset the BI-RADS ratings are not available but each breast has been confirmed by biopsy to be benign or malignant, so these binary labels were used.

3.5 Results

In this section, overall performance of each experiment on all test sets is first reported in Subsection 3.5.1, based on the breast-wise AUC-ROC scores in Table 3.2. Secondly, in Subsection 3.5.2 the sensitivity and specificity resulting from the various treatments is analysed, as captured by Table 3.3.

TABLE 3.2: AUC (±standard error) for each combination of training data and test data, for the various experimental treatments. Shaded are results where the training and test sets are from the same distribution.

Training Dataset	Treatment	Test Dataset			
		HMI	VinDR	CMMD	INBreast
HMI	Imbalanced	0.776 (±0.021)	0.700 (±0.032)	0.646 (±0.029)	0.818 (±0.033)
	Under-sampled	0.767 (±0.023)	0.664 (±0.027)	0.654 (±0.030)	0.803 (±0.040)
	Over-sampled	0.762 (±0.023)	0.699 (±0.028)	0.641 (±0.031)	0.789 (±0.039)
	Class Weighting	0.786 (±0.021)	0.650 (±0.030)	0.660 (±0.030)	0.795 (±0.037)
	Artifacted	**0.807** (±0.021)	**0.760** (±0.025)	**0.730** (±0.028)	**0.845** (±0.030)
VinDR	Imbalanced	0.622 (±0.026)	0.757 (±0.028)	0.671 (±0.027)	0.732 (±0.045)
	Under-sampled	0.630 (±0.026)	0.691 (±0.027)	0.680 (±0.030)	0.744 (±0.045)
	Over-sampled	0.646 (±0.028)	0.752 (±0.027)	0.668 (±0.030)	**0.824** (±0.038)
	Class Weighting	0.589 (±0.027)	0.739 (±0.030)	**0.685** (±0.029)	0.691 (±0.048)
	Artifacted	**0.686** (±0.025)	**0.768** (±0.027)	0.644 (±0.030)	0.799 (±0.039)
CMMD	Imbalanced	0.516 (±0.027)	0.520 (±0.032)	**0.727** (±0.026)	0.656 (±0.049)
	Under-sampled	0.559 (±0.028)	0.513 (±0.032)	0.690 (±0.028)	0.702 (±0.045)
	Over-sampled	**0.614** (±0.026)	**0.667** (±0.031)	0.719 (±0.028)	**0.711** (±0.048)
	Class Weighting	0.495 (±0.027)	0.541 (±0.027)	0.681 (±0.027)	0.624 (±0.048)

3.5.1 AUC Results

When training on $\mathcal{D}_{\text{HMI}_{\text{train}}}$ and testing on the corresponding test set $\mathcal{D}_{\text{HMI}_{\text{test}}}$, there are no significant differences between simply training on the imbalanced dataset (AUC=0.776) and applying standard class imbalance techniques. Whereas the class weighting leads to a slight increase in AUC (Δ=0.01), over-sampling is associated with a decrease of 0.014 to 0.762. There is a wider range of results when considering out-of-distribution generalisation. Of these four initial experiments, the imbalanced training leads to the best performance on $\mathcal{D}_{\text{VinDR}_{\text{test}}}$ and INBreast, whereas the models trained with class weighting perform slightly better on $\mathcal{D}_{\text{CMMD}_{\text{test}}}$ (0.660 vs. 0.646), although the generalisation to this dataset remains poor.

The models trained on $\mathcal{D}_{\text{HMI}_{\text{train}}}$ using the synthetic lesions (*Artifacted*) method achieve the best results on the related $\mathcal{D}_{\text{HMI}_{\text{test}}}$, with an improvement of Δ=0.021 over the next best AUC score. Moreover, this method achieves a noticeable increase in generalisability to out-of-distribution test sets, with improvements over the next best result for $\mathcal{D}_{\text{VinDR}_{\text{test}}}$ (Δ=0.060), $\mathcal{D}_{\text{CMMD}_{\text{test}}}$ (Δ=0.070), and INBreast (Δ=0.027). Most notably, the resulting generalisability leads to achieving the same performance on out-of-distribution datasets as when training directly on the regular images from those distributions ($\mathcal{D}_{\text{VinDR}_{\text{test}}}$: 0.760 vs. 0.757, $\mathcal{D}_{\text{CMMD}_{\text{test}}}$: 0.730 vs. 0.727).

Evaluating the models trained on $\mathcal{D}_{\text{VinDR}_{\text{train}}}$, there are minor drops in the AUC on $\mathcal{D}_{\text{VinDR}_{\text{test}}}$ when comparing the imbalanced training (0.757) to over-sampling (0.752), and class weighting (0.739). However, a larger decrease occurs when under-sampling (0.691, Δ=-0.066), where a large portion of the benign samples are removed from the training set. The over-sampled model achieves the best results on INBreast (0.824), and $\mathcal{D}_{\text{HMI}_{\text{test}}}$ (0.646), although this latter result indicates that the models trained on VinDR data fail to generalise well to HMI images in general.

The *Artifacted* method, trained on $\mathcal{D}_{\text{VinDR}_{\text{train}}}$, demonstrates an improvement in AUC over the next best method on $\mathcal{D}_{\text{VinDR}_{\text{test}}}$, i.e., the imbalanced training (Δ=0.011), as well as better generalisability to $\mathcal{D}_{\text{HMI}_{\text{test}}}$ (Δ=0.040). It also yields the second best AUC score on INBreast (0.799) among the five results in this second set of experiments. In contrast, the *Artifacted* method with $\mathcal{D}_{\text{VinDR}_{\text{train}}}$ delivers the lowest result when generalising to $\mathcal{D}_{\text{CMMD}_{\text{test}}}$, with a difference of Δ=-0.041 compared to class weighting.

The models in the final set of experiments are trained on $\mathcal{D}_{\text{CMMD}_{\text{train}}}$, and in this case, there is no *Artifacted* experiment as the nature of the imbalance is different in the CMMD dataset, i.e., there are more malignant samples than benign (see Table 3.1). The imbalanced (0.727) and over-sampled (0.719) experiments offer similar AUC performance when tested on $\mathcal{D}_{\text{CMMD}_{\text{test}}}$, with a sizeable gap over the next two results, under-sampled (0.690) and class weighting (0.681). Most of the trained models perform very poorly when applied to $\mathcal{D}_{\text{HMI}_{\text{test}}}$ and $\mathcal{D}_{\text{VinDR}_{\text{test}}}$, with AUC scores close to 0.5, which would be achieved by a predictor making random guesses. The over-sampling experiment is an exception to this with higher AUC scores for $\mathcal{D}_{\text{HMI}_{\text{test}}}$ (0.614), $\mathcal{D}_{\text{VinDR}_{\text{test}}}$ (0.667), and INBreast (0.711). These results are still poor, however, and indeed the CMMD dataset is associated with the lowest results across all experiments, both in terms of evaluating directly on $\mathcal{D}_{\text{CMMD}_{\text{test}}}$, and the ability of a model trained on $\mathcal{D}_{\text{CMMD}_{\text{train}}}$ to generalise to any of the other three datasets.

3.5.2 Sensitivity & Specificity

The first result that is evident from the sensitivity and specificity results reported in Table 3.3 is that an imbalance in the training data is reflected in imbalances between

TABLE 3.3: Metrics (±standard error) where the test set and training set come from the same distribution.

Training Dataset	Treatment	Sensitivity	Specificity	MCC
HMI	Imbalanced	0.574 (±0.033)	0.824 (±0.027)	0.418 (±0.042)
	Under-sampled	0.690 (±0.034)	0.709 (±0.029)	0.399 (±0.043)
	Over-sampled	0.595 (±0.033)	0.776 (±0.031)	0.379 (±0.037)
	Class Weighting	0.515 (±0.034)	**0.857** (±0.023)	0.398 (±0.044)
	Artifacted	**0.753** (±0.028)	0.689 (±0.031)	**0.443** (±0.043)
VinDR	Imbalanced	0.004 (±0.010)	1.000 (±0.000)	0.033 (±0.063)
	Under-sampled	0.400 (±0.048)	0.837 (±0.008)	0.141 (±0.032)
	Over-sampled	0.263 (±0.046)	**0.989** (±0.004)	**0.378** (±0.046)
	Class Weighting	**0.644** (±0.043)	0.708 (±0.011)	0.170 (±0.022)
	Artifacted	0.499 (±0.052)	0.873 (±0.009)	0.253 (±0.030)
CMMD	Imbalanced	0.781 (±0.027)	0.479 (±0.047)	**0.261** (±0.052)
	Under-sampled	0.336 (±0.028)	**0.894** (±0.029)	0.236 (±0.040)
	Over-sampled	**0.803** (±0.026)	0.433 (±0.047)	0.246 (±0.055)
	Class Weighting	0.520 (±0.030)	0.727 (±0.032)	0.245 (±0.041)

sensitivity and specificity in the trained models. 70% of training samples in $\mathcal{D}_{\text{HMI}}$ are benign, resulting in a lower sensitivity (Se) than specificity (Sp) in the trained model (Se=0.574, Sp=0.824). Similarly, 70% of the images in $\mathcal{D}_{\text{CMMD}}$ are malignant, resulting in a higher sensitivity (Se=0.781, Sp=0.479). The most severe example of this phenomenon is visible with $\mathcal{D}_{\text{VinDR}}$, where benign samples outnumber malignant 19:1, and as a result nearly all test samples are classified as benign by the trained model (Se=0.004, Sp=1.0). In general, the techniques employed to tackle class imbalance reduced this disparity between sensitivity and specificity. The only exceptions to this were the application of class weighting with $\mathcal{D}_{\text{HMI}}$, where sensitivity reduced from the imbalanced case, and over-sampling with $\mathcal{D}_{\text{CMMD}}$, which saw a decrease in specificity.

Despite the imbalance between sensitivity and specificity when training on the imbalanced datasets, the results for MCC are in fact higher than when applying the three standard class imbalance tecniques (under-sampling, over-sampling, and class weighting) for both $\mathcal{D}_{\text{HMI}}$ and $\mathcal{D}_{\text{CMMD}}$. This observation does not hold when the class imbalance is more severe with $\mathcal{D}_{\text{VinDR}}$, where the near-zero sensitivity contributes to a low MCC of 0.033.

The *Artifacted* method achieves the highest MCC score for $\mathcal{D}_{\text{HMI}}$ (MCC=0.443), as well as the second best result for $\mathcal{D}_{\text{VinDR}}$ (0.253). The highest result for the latter case is given by the oversampled experiment (0.378); however, this coincides with a low sensitivity (0.263), which indicates that the model may have difficulty generalising to malignant samples different from those in the training set.

Figure 3.2 demonstrates the behaviour of the models trained on $\mathcal{D}_{\text{VinDR}_{\text{train}}}$ under different treatments. When trained on the imbalanced dataset, the model has a strong bias towards the benign class, meaning that essentially every sample is predicted to have a low probability of malignancy at test-time, resulting in the low sensitivity and high specificity seen at an operating point of 0.5. However, the trained model remains capable of achieving some separation between malignant and benign samples, leading to the second highest AUC score among the VinDR experiments. For example, over 70% of truly benign samples gain a predicted malignancy score between 0-0.05, whereas less than 40% of truly malignant are placed in this bracket.

Over-sampling resolves the problem of every sample being classified as benign. However, the resulting behaviour remains unusual and undesirable. Nearly 100% of

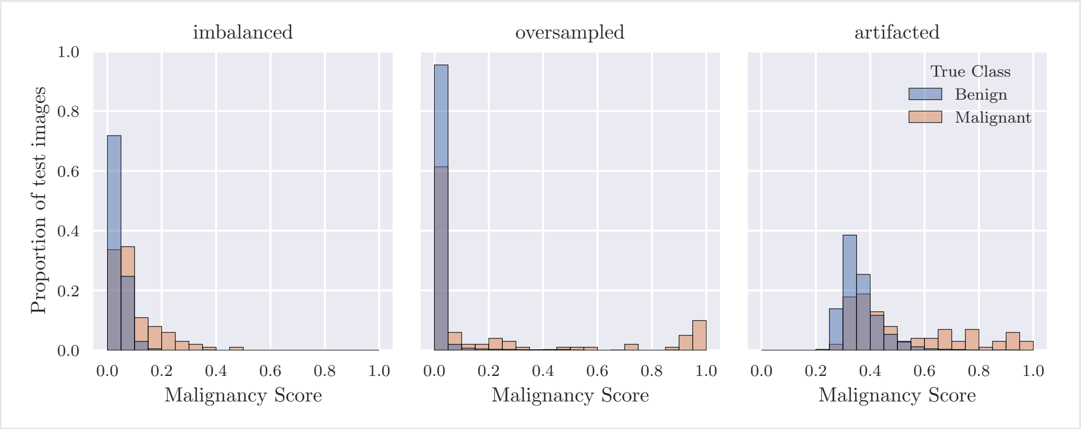

FIGURE 3.2: Distributions of predictions for selected models, trained on $\mathcal{D}_{\text{VinDR}_{\text{train}}}$ and applied to $\mathcal{D}_{\text{VinDR}_{\text{test}}}$. Malignancy score is the softmax output of the model - at an operating point of 0.5, for example, the model would predict malignant for samples above 0.5, and benign for those below. The distributions in the plot are normalised so that bar heights sum to 1 for each class in each plot.

normal or benign breasts, as well as over 60% of breasts with malignant lesions, are assigned a low score of malignancy between 0-0.05. Conversely, almost 20% of malignant samples attain a high malignancy score between 0.9-1. This behaviour may indicate overfitting on the small number of malignant samples in the training set, leading the model to correctly predict malignancy with high confidence for samples who share some characteristics with those in the training set, while predicting most samples in the test set to be benign with high confidence.

The *Artifacted* model yields a more natural distribution of predictions on the test set in that the range is more spread, allowing high confidence predictions to be separated more easily from low confidence predictions. However, two problems are evident in the third histogram in Figure 3.2. Firstly, a large proportion of truly malignant breasts still attain predicted malignancy scores similar to those assigned to benign samples, particularly in the range 0.3-0.45. Secondly, virtually zero samples are assigned a malignancy score less than 0.25. The exact reason for this behaviour remains unclear, but potential causes and solutions are discussed in Section 3.6.

3.6 Discussion

RQ1 – Class Imbalance Effect

For each of the three imbalanced datasets, the class imbalance caused the classifier to bias towards the majority class. As might be expected, a higher imbalance resulted in a higher bias, i.e., a standard classifier trained on the VinDR dataset predicted every test sample to be benign. On the other hand, models trained on imbalanced datasets unexpectedly achieved comparable AUC-ROC scores to models trained with common class imbalance techniques, indicating that despite predictions shifting towards the majority class, the model still learns to separate the two classes. This lends credence to setting a new threshold on the output scores post-training as an effective technique of dealing with class imbalance (Buda, Maki, and Mazurowski, 2018). New thresholds can also be set in practice, regardless of class imbalance, to achieve a certain balance between sensitivity and specificity. Therefore, threshold setting may be the simplest approach to tackling class imbalance, while remaining relatively effective.

RQ2 – Common Techniques

Among over-sampling, under-sampling, and class weighting, no single technique consistently achieved the best AUC-ROC nor MCC scores across all experiments. Over-sampling performed best of these techniques on the VinDR and CMMD experiments, in the latter case also leading to a significant improvement when testing on out-of-distribution datasets. However, over-sampling performed worst on three out of four test sets when trained on HMI. Therefore, there is clearly no one single best approach for every mammography dataset. However, one useful lesson is that experimenting with different techniques may allow for an improved out-of-distribution generalisation without sacrificing in-distribution performance, as in the case of over-sampling on CMMD.

There is a notable drop in classification performance in the VinDR and CMMD experiments when under-sampling is applied. This aligns with expectations that using only a subset of data may remove informative samples and thus may hinder performance, particularly in the case of VinDR where 90% of the dataset is removed. Furthermore, the model began to learn only after several attempts with different random seeds for the initialisation of model weights and taking a different random sample of benign images. The training process was not straightforward either for the imbalanced experiment nor class weighting on the VinDR dataset, which may have been due to the low frequency with which malignant samples were seen during training, and taking a different random order of the training data was necessary for the model to begin learning.

Over-sampling, together with the *Artifacted* method, were the only VinDR experiments that worked without further adjustment. Moreover, the model trained on VinDR with over-sampling performed well on the test set relative to the other methods according to the AUC-ROC and MCC metrics, so over-sampling could be seen as a promising approach for cases with high imbalance between classes. On the other hand, the distribution of predictions of this model showed signs of overfitting on the small number of minority samples in the training set, which is a common concern of using over-sampling. This output distribution may be problematic in cases where the softmax outputs are used to gain information about the prediction uncertainty, e.g., in Tardy, Scheffer, and Mateus (2019). Future research is required to determine whether this behaviour can be resolved by, for example, further data augmentation or a combined over-sampling/under-sampling approach such as that used by Buda, Maki, and Mazurowski (2018).

RQ3 – Synthetic Lesions

The results showed that inserting synthetic lesions into benign samples during training can be a useful technique to balance classes. When applied to the HMI data, the *Artifacted* method achieves an improvement of 0.021 in ROC-AUC over the next best method, as well as significant improvements in ROC-AUC on the out-of-distribution test sets, up to Δ=0.07. Applied to the more highly imbalanced data, VinDR, the performance improvement is more modest (Δ=0.011), and the out-of-distribution performance is mixed. Despite these mixed results, the *Artifacted* model shows more promise as it does not suffer from the overfitting behaviour demonstrated by the model trained with over-sampling.

The distribution of output predictions for the *Artifacted* model (Figure 3.2) shows that no samples were given a malignancy score less than 0.25 (where if a sample receives a low score, the model is more confident it is benign). The cause of this is unclear, but perhaps either noisy labels during training, or difficulty in distinguishing malignant lesions from benign lesions may lead to this behaviour. Firstly, with regard to noisy labels, for approximately 10% of malignant breasts in the VinDR dataset, malignant lesions appear in only one of the two views (CC or MLO), and so the other image will be included as malignant during training despite containing no signs of malignancy. Predicting these images to be benign during training would lead to a high contribution to the loss, perhaps causing the model to avoid classifying images as benign with high confidence. Secondly, benign masses and calcifications can appear in breasts in the non-cancerous group of images. If the model, for example, learns during training that masses in general are correlated with malignancy, it may incorrectly predict a non-cancerous breast containing a benign mass to be malignant. Again, this would lead to a large contribution to the loss, potentially preventing the model from predicting samples to be benign with high confidence. If this were the case, improved results might be gained by a two-model process where the first model classifies normal (i.e., no lesions) from abnormal breasts, and the second model determines whether the detected lesion is benign or malignant.

Comparison to previous work

Our findings both support and disagree with some of the findings of Buda, Maki, and Mazurowski (2018) who carried out a systematic review of the behaviour of CNNs for digit recognition under class imbalance. That study found that over-sampling generally performed best in terms of AUC-ROC results, but in fact the performance diverged from the baseline only as the number of minority classes and the imbalance ratio increased. In the current study, there is only one minority class, and smaller imbalance ratios (at most 19:1) than many of the situations they examined, so this aligns with the current results which show that most standard techniques for dealing with class imbalance provided no significant benefit to AUC-ROC over the baseline. The authors also claimed that over-sampling with CNNs did not cause overfitting. However, the overconfidence of the model trained on the VinDR dataset with oversampling in the current study indicates that overfitting may still remain a problem.

The aim of this study was not to achieve state-of-the-art performance, but rather to take a relatively standard methodology and conduct a fair and thorough comparison of common techniques used to tackle class imbalance in mammography. It is therefore no surprise that the results achieve lower performance than current state-of-the-art methods, which focus on optimising a single method on a single dataset to achieve the best results possible.

Stadnick et al. (2021) test several state-of-the-art models on multiple datasets, including INBreast and CMMD. However, they use only 28% of the INBreast data in their test set, and test on the full CMMD dataset, so the results are not directly comparable to this study, but may provide an indication of performance. The reported AUC-ROC scores vary between 0.612-0.980 on INBreast and 0.449-0.831 for CMMD. The *Artifacted* method achieves an AUC-ROC of 0.845 on INBreast and 0.730 on CMMD, comparable to the model of Wu et al. (2019), which scores 0.802 and 0.740, respectively (Stadnick et al., 2021).

The VinDR and CMMD datasets are both relatively new, and so are not as well studied as INBreast or other older public mammography datasets. Thus, this study provides a useful reference and baseline for other researchers using these datasets. Generally low results were achieved on the CMMD dataset, both in this study and found by Stadnick et al. (2021), and one potential cause of this is the composition of the dataset. All breasts in the CMMD dataset have been biopsied, indicating that there are suspicious abnormalities in both the malignant and benign images, making the task of separating malignant from non-malignant more difficult. Wu et al. (2019) conducted experiments on a different dataset, and found lower classification performance on a biopsied sub-population compared to the overall screening population.

Implications and Recommendations

Why is class imbalance detrimental? Because it biases trained models towards predicting the majority class. The results show that despite this shift in output prediction scores towards the majority class, however, the model can still learn to separate malignant from benign samples as much as when typical class imbalance techniques are applied, as evidenced by the similar AUC-ROC scores. Thus, if the imbalance ratio is not extremely high, setting a (data-specific) threshold on the softmax outputs should suffice to achieve the desired trade-off between sensitivity and specificity. On the other hand, if the context necessitates more realistic predicted probabilities of malignancy, indicative of the confidence of the predictions for individual samples, then applying a technique for dealing with this imbalance will be important.

For higher class imbalances, the experiment with VinDR indicates that oversampling may provide the best separation between malignant and benign, but the output distributions of the predictions should be assessed to ensure the model does not overfit on the small number of minority samples. Finally, if feasible within the context of the study, whether on mammography or medical imaging more broadly, generating synthetic lesions could provide a good way of balancing the classes while also introducing prior knowledge from domain experts, and potentially improving both in-distribution and out-of-distribution generalisation.

Limitations

Due to the large number of experiments, and the training times for each training run of 1-8 days, it was not feasible to conduct extensive hyperparameter tuning. While the application of techniques such as Batch Normalisation and Instance Normalisation have been shown to increase robustness with respect to hyperparameters, better results may have been possible for each experiment by finding the best hyperparameters in each case. Instead, the same treatment is given to all experiments to allow a fair comparison of many methods within a reasonable time frame.

There are many proposed techniques for tackling class imbalance, and only a small subset of these are considered here, i.e., the most common methods. However, it is possible that other, and in particular, more complex, methods may improve results similar to Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) improving performance for traditional machine learning problems. More complex methods are, by their nature, more difficult to implement, which means that it may take time for a single technique to become more popular than simpler sampling and weighting strategies. Until such time as one of the many techniques available proves superior for many domains and tasks, this study will provide a useful reference to researchers in mammography, and medical imaging more broadly.

Chapter 4

Evaluation

4.1 Internship Review

The internship was a highlight of the two year master's programme, and I believe that I achieved the goals I set out for myself, i.e., to do something useful while gaining experience in research, in a start-up, and in AI for medical imaging. The six months at Hera-MI gave me a view of the team dynamic and camaraderie across the whole company, while the small size of the company provided a stronger sense of connection to the overall activities of the company. I learned about the development of commercial solutions based on AI for medical imaging, and the inherent challenges and opportunities. Carrying out wide ranging experiments and staying up-to-date with the latest works helped me to improve my research skills, which I will further hone over the next three years during my PhD. My final goal was to do work that is useful, i.e., make a meaningful contribution to a worthwhile industry and have an impact on company activity, and in my opinion I also achieved this.

The impact that I had on the company came both from my research outputs, as well as the development of the techniques used during the course of this research. The conclusions from the class imbalance study inform future work in several ways. Firstly, when using datasets with a minor imbalance, researchers in the company may use the imbalanced datasets, without spending time on implementing methods to tackle the imbalance. Secondly, my results on the more recent datasets (CMMD and VinDR) can serve as a baseline and reference for research on these datasets, and similarly with the outputs of the exploratory analysis on these datasets. Finally, building on the results of this study, the artifacted method for classification problems will be further investigated, and integrated into models for new releases of BreastSlimView if it continues to show promising results.

My other contribution to the company was in developing code in the shared codebase. My implementation of the class imbalance techniques was directly integrated into a generalised model training framework within the company, and other team members can easily use these methods in the future in a flexible manner. Other utility functions that I created will also be useful for the team, including a function converting detailed JSON files of image annotations data into Pandas DataFrames for easy manipulation, as well as code I wrote for various data visualisation tasks, e.g., intensity distribution comparisons and plots of a model's latent space with t-SNE representations, which will be useful to the team for future data exploration, quality documentation, and when writing papers.

Overall, I believe that I worked well with the team, building positive relationships with my supervisor, the other members of the research team, and beyond. In general, I worked with a high degree of autonomy, investigating problems and potential solutions before asking the advice of others. However, in hindsight, there remains some room for improvement in this aspect of my working style, as at times I worked too autonomously, perhaps spending an hour investigating how a particular piece of code works when asking someone else may have been much more efficient.

4.2 Technical Lessons

Although I had worked on several deep learning projects before starting the internship, the large number of experiments that I conducted brought up problems that I had not encountered before, and which I resolved by exploring and by discussing with the rest of the team, who may have had experienced similar problems. For example, for one experiment the model learned nothing for several epochs and then started learning normally, and a colleague suggested that the initial random weights might be sub-optimal, and so simply setting a different random seed and re-initialising the model resolved this behaviour.

For several experiments, the order in which the data appeared during each epoch had a significant effect on the results. For example, seeing mostly benign images for the first half of an epoch and then balanced data for the second half prevented the model from learning in one case (i.e., AUC-ROC on the validation remained close to 0.5), and simply shuffling the data resolved this. Similarly, when training with synthetic lesions, if the real malignant images were all processed first in an epoch and the synthetic lesions came only at the end of an epoch, the resulting model performed worse than those trained with other class imbalance techniques. Ensuring that both real and synthetic malignant lesions occur together in each batch led to the higher results shown in Section 3.5. Indeed, the order of the data during training was not something to which I had previously given much thought, but I wish to investigate this further in future research, as well as other ordering techniques (e.g., curriculum learning).

During most of my university projects, I worked on Jupyter Notebooks or Google Colab, managing a single relatively simple model on one dataset, and so conducting experiments on a larger scale in the internship certainly presented some new challenges and lessons. I developed code to run in multiple environments, i.e., locally for development and de-bugging, as well as on three different servers for various purposes. Therefore, setting up isolated environments with **virtual environments** in Python proved very useful to manage package dependencies, and quickly set up a new environment when needed. Working with a **Git** version control system was important in ensuring the code running on each system was up-to-date and was the same in all cases. This was also essential for collaboration on code with the rest of the team.

I improved my coding practices based on observing the more experienced members of my team. Investing time at the start to set up a good project framework saved countless hours later on in the project. For example, making my experiments **configurable** by providing arguments in files rather than hard-coding them into code, allowed me to be quickly change parameters between experiments and store them for future reference. Furthermore, informative **logs** were important to ensure the training was progressing as expected or to enable quick de-bugging if not, and tracking the evolution of training using **Tensorboard** allowed me to observe when an experiment was not working, and change it, rather than waiting several days for the training to finish.

Much of my time was spent on developing code in Python, which I had been using throughout my master's programme, and the internship provided further opportunity to develop my skills. On the advice of other team members, I began using the PyCharm Integrated Development Environment (IDE) for developing code, and I found the real-time error and warning flags particularly useful. Another recommendation was Conda for installing Python packages, and I found this especially convenient for installing Tensorflow and its various dependencies, as otherwise it can be tedious to find and install the correct dependencies. I also improved my knowledge of Linux and writing bash scripts for running code on servers. Finally, I used Fiji ImageJ for viewing and manual adjustment of images (such as increasing contrast), which will prove useful in my future research on medical images.

4.3 Personal Experience

4.3.1 Internship in a Foreign Country

Hera-MI is a French company with international aspirations, and it was interesting to see how this manifested in the day-to-day activities fo the company. For example, most documentation is prepared, as standard, in both English and French, and the presentation slides of the research team are also prepared in English. I was one of only two non-French people working in the Hera-MI office in Nantes, however, so daily communication was of course through French, although I spoke English during one-on-one meetings with my supervisor, and when presenting at weekly meetings. Given that my level of spoken French is not quite adequate, I had known prior to commencing the internship that this would present challenges, but I believed that this was one of the best ways to improve – to "dive in at the deep end", so to speak. Indeed, I believe language barriers hindered my full integration into the office culture, as I often struggled to keep up with the pace of group conversations. On the other hand, the team were welcoming and patient with me, and we shared many friendly moments together. Moreover, my level of French has certainly improved over these six months, as well as my knowledge of and appreciation for French culture, so the extra challenge proved worthwhile.

4.3.2 Reflections

My experience with deep learning projects during university was useful preparation for the internship at Hera-MI. One difference between these two activities was immediately evident: the synergy of many researchers working together on similar tasks. Each university project involved coding from scratch for everything from data handling, pre-processing and modelling, whereas Hera-MI has built up a solid codebase over several years. As the team grows, this is becoming more structured, better documented, and more diverse, allowing team members to easily move and clean raw images, utilise ready-to-use deep learning model architectures, and perform standardised evaluation of results. The company has also curated many private and public datasets, processing and storing them in a standardised way to make access and analysis easier. The accumulation of these efforts allows each researcher to work more efficiently, and the efficiency gains will grow as the team and codebase develop.

I worked in data analytics for four years in a Fast-Moving Consumer Goods (FMCG) company before commencing the master's degree, and so it was interesting for me to compare my experience during the internship to my previous work. The difference stood out to me in three aspects in particular: 1) the length of research cycles, 2) working within a specialised team at the heart of the business, and 3) the awareness of the uncertain nature of machine learning projects.

Firstly, research in AI takes time. Although during the latter stages of a project, it may be possible to quickly iterate through several versions of the developed model, at the start of a project it is necessary to thoroughly plan experiments, investigating the data to be used, setting up the model and training framework, etc. Furthermore, in medical contexts patient safety is paramount, so everything that is implemented must be thoroughly evaluated before deployment. This was different to some of the previous teams I worked with, particularly during a period working in the e-commerce part of my previous company, where new small-scale projects could be conceived and concluded within the same week.

Secondly, whereas my previous role as a data analyst was to support the main business activities by providing insights to help decision-making, the main product of Hera-MI is built on data and machine learning. This means that there is a large research and development team whose members can communicate effectively with each other using technical language. If I presented a graph at the weekly meeting, for example, and stated simply that my model was ’overfitting‘, this would be understood without further elaboration or explanation, and moreover other members of the team might offer ways to deal with the problem. In this way, working on a specialised team like this enabled more fluid communication and a quicker development of deep technical skills.

Finally, again because the company is built around data and machine learning, there is an understanding in the team and broader company of the inherent uncertainty of machine learning projects. For a data analytics project where the objective is to answer a question from data, or to create a report on historical data, as long as the data is available and relatively accurate, there is a strong chance the project will be successful, i.e., the question will be answered or or the report built. In contrast, the idea of success in a predictive machine learning project is to predict phenomena with a certain level of accuracy, which simply may not be possible at the time given the available data and techniques. This is something that is well understood in companies like Hera-MI, and to which more traditional businesses who experiment with machine learning must learn to adapt. Some attempts will not work, but some will, and so the old proverb rings true: ”If at first you don’t succeed, try, try, and try again“.

Bibliography

Abdelhafiz, Dina et al. (2019). "Deep convolutional neural networks for mammography: Advances, challenges and applications". *BMC Bioinformatics* 20.11, pp. 1–20. DOI: 10.1186/S12859-019-2823-4/FIGURES/3.

Ali Madani, Tanveer et al. (2018). "Chest x-ray generation and data augmentation for cardiovascular abnormality classification". *Medical imaging 2018: Image processing* 10574.2, pp. 415–420. DOI: 10.1117/12.2293971.

Alyafi, Basel, Oliver Diaz, and Robert Martí (2020). "DCGANs for realistic breast mass augmentation in x-ray mammography". *Medical Imaging 2020: Computer-Aided Diagnosis*. Ed. by Horst K. Hahn and Maciej A. Mazurowski. SPIE, p. 68. DOI: 10.1117/12.2543506.

Anand, Rangachari et al. (1993). "An Improved Algorithm for Neural Network Classification of Imbalanced Training Sets". *IEEE Transactions on Neural Networks* 4.6, pp. 962–969. DOI: 10.1109/72.286891.

Bahl, Manisha et al. (2015). "Architectural Distortion on Mammography: Correlation With Pathologic Outcomes and Predictors of Malignancy". *American Journal of Roentgenology* 205.6, pp. 1339–1345. DOI: 10.2214/AJR.15.14628.

Bakalo, Ran, Rami Ben-Ari, and Jacob Goldberger (2019). "Classification and detection in mammograms with weak supervision via dual branch deep neural net". *Proceedings - International Symposium on Biomedical Imaging*. Vol. 2019-April. IEEE Computer Society, pp. 1905–1909. DOI: 10.1109/ISBI.2019.8759458.

Bevers, Therese B. et al. (2018). "Breast Cancer Screening and Diagnosis, Version 3.2018, NCCN Clinical Practice Guidelines in Oncology". *Journal of the National Comprehensive Cancer Network* 16.11, pp. 1362–1389. DOI: 10.6004/jnccn.2018.0083.

Boughorbel, Sabri, Fethi Jarray, and Mohammed El-Anbari (2017). "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric". *PLoS ONE* 12.6, e0177678. DOI: 10.1371/journal.pone.0177678.

Bowles, Christopher et al. (2018). "GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks". *arXiv preprint*. DOI: 10.48550/arXiv.1810.10863.

Bradley, Andrew P. (1997). "The use of the area under the ROC curve in the evaluation of machine learning algorithms". *Pattern Recognition* 30.7, pp. 1145–1159. DOI: 10.1016/S0031-3203(96)00142-2.

Brady, Adrian P. (2017). "Error and discrepancy in radiology: inevitable or avoidable?" *Insights into Imaging* 8.1, pp. 171–182. DOI: 10.1007/S13244-016-0534-1/TABLES/3.

Bria, Alessandro, Claudio Marrocco, and Francesco Tortorella (2020). "Addressing class imbalance in deep learning for small lesion detection on medical images". *Computers in Biology and Medicine* 120, p. 103735. DOI: 10.1016/J.COMPBIOMED.2020.103735.

Broeders, Mireille et al. (2012). "The impact of mammographic screening on breast cancer mortality in Europe: A review of observational studies". *Journal of Medical Screening* 19.SUPPL. 1, pp. 14–25. DOI: 10.1258/jms.2012.012078.

Buda, Mateusz, Atsuto Maki, and Maciej A. Mazurowski (2018). "A systematic study of the class imbalance problem in convolutional neural networks". *Neural Networks* 106, pp. 249–259. DOI: 10.1016/J.NEUNET.2018.07.011.

Cha, Kenny H. et al. (2019). "Evaluation of data augmentation via synthetic images for improved breast mass detection on mammograms using deep learning". *Journal of Medical Imaging* 7.01, p. 1. DOI: 10.1117/1.jmi.7.1.012703.

Chawla, Nitesh V., Nathalie Japkowicz, and Aleksander Kotcz (2004). "Editorial: Special Issue on Learning from Imbalanced Data Sets". *ACM SIGKDD Explorations Newsletter*. Vol. 6. 1, pp. 1–6. DOI: 10.1145/1007730.1007733.

Chawla, Nitesh V. et al. (2002). "SMOTE: Synthetic Minority Over-sampling Technique". *Journal of Artificial Intelligence Research* 16, pp. 321–357. DOI: 10.1613/JAIR.953.

Cheng, Chi-Tung et al. (2021). "A scalable physician-level deep learning algorithm detects universal trauma on pelvic radiographs". *Nature Communications* 12.1, p. 1066. DOI: 10.1038/s41467-021-21311-3.

Chicco, Davide and Giuseppe Jurman (2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". *BMC Genomics* 21.1. DOI: 10.1186/s12864-019-6413-7.

Choukroun, Yoni et al. (2017). "Mammogram Classification and Abnormality Detection from Nonlocal Labels using Deep Multiple Instance Neural Network". *Eurographics Proceedings*, pp. 11–19. DOI: 10.2312/VCBM.20171232.

Conant, Emily F. et al. (2019). "Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis". *Radiology: Artificial Intelligence* 1.4, e180096. DOI: 10.1148/ryai.2019180096.

Cui, Chunyan et al. (2021). "The Chinese Mammography Database (CMMD): An online mammography database with biopsy confirmed types for machine diagnosis of breast." *The Cancer Imaging Archive* 1. DOI: 10.7937/tcia.eqde-4b16.

Dildar, Mehwish et al. (2021). "Skin Cancer Detection: A Review Using Deep Learning Techniques". *International Journal of Environmental Research and Public Health* 18.10, p. 5479. DOI: 10.3390/ijerph18105479.

Doi, Kunio (2007). "Computer-aided diagnosis in medical imaging: Historical review, current status and future potential". *Computerized Medical Imaging and Graphics* 31.4-5, pp. 198–211. DOI: 10.1016/j.compmedimag.2007.02.002.

Efron, Bradley and Robert J. Tibshirani (1993). *An introduction to the bootstrap*. Boca Raton: Chapman & Hall, pp. 45–82.

Frid-Adar, Maayan et al. (2018). "Synthetic data augmentation using GAN for improved liver lesion classification". *Proceedings - International Symposium on Biomedical Imaging* 2018-April, pp. 289–293. DOI: 10.1109/ISBI.2018.8363576.

Geller, Berta M. et al. (2009). "Radiologists' Performance and Their Enjoyment of Interpreting Screening Mammograms". *AJR. American journal of roentgenology* 192.2, p. 361. DOI: 10.2214/AJR.08.1647.

Japkowicz, Nathalie and Shaju Stephen (2002). "The class imbalance problem: A systematic study". *Intelligent Data Analysis* 6.5, pp. 429–449. DOI: 10.3233/IDA-2002-6504.

Johnson, Justin M. and Taghi M. Khoshgoftaar (2019). "Survey on deep learning with class imbalance". *Journal of Big Data* 6.1, pp. 1–54. DOI: 10.1186/S40537-019-0192-5/TABLES/18.

Korkinof, Dimitrios et al. (2018). "High-Resolution Mammogram Synthesis using Progressive Generative Adversarial Networks". *arXiv preprint* arXiv:1807.03401.

Krawczyk, Bartosz (2016). "Learning from imbalanced data: open challenges and future directions". *Progress in Artificial Intelligence* 5.4, pp. 221–232. DOI: 10.1007/S13748-016-0094-0/TABLES/1.

Kriege, Mieke et al. (2004). "Efficacy of MRI and Mammography for Breast-Cancer Screening in Women with a Familial or Genetic Predisposition". *New England Journal of Medicine* 351.5, pp. 427–437. DOI: 10.1056/NEJMoa031759.

Lehman, Constance D. et al. (2015). "Diagnostic accuracy of digital screening mammography with and without computer-aided detection". *JAMA Internal Medicine* 175.11, pp. 1828–1837. DOI: 10.1001/jamainternmed.2015.5231.

Li, Zeju, Konstantinos Kamnitsas, and Ben Glocker (2021). "Analyzing Overfitting under Class Imbalance in Neural Networks for Image Segmentation". *IEEE Transactions on Medical Imaging* 40.3, pp. 1065–1077. DOI: 10.1109/TMI.2020.3046692.

Lundervold, Alexander Selvikvåg and Arvid Lundervold (2019). "An overview of deep learning in medical imaging focusing on MRI". *Zeitschrift für Medizinische Physik* 29.2, pp. 102–127. DOI: 10.1016/j.zemedi.2018.11.002.

Mandelblatt, Jeanne S. et al. (2016). "Collaborative modeling of the benefits and harms associated with different U.S. Breast cancer screening strategies". *Annals of Internal Medicine* 164.4, pp. 215–225. DOI: 10.7326/M15-1536.

Mazurowski, Maciej A. et al. (2008). "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance". *Neural Networks* 21.2-3, pp. 427–436. DOI: 10.1016/J.NEUNET.2007.12.031.

McKinney, Scott Mayer et al. (2020). "International evaluation of an AI system for breast cancer screening". *Nature* 577.7788, pp. 89–94. DOI: 10.1038/s41586-019-1799-6.

Moreira, Inês C. et al. (2012). "INbreast: Toward a Full-field Digital Mammographic Database." *Academic Radiology* 19.2, pp. 236–248. DOI: 10.1016/j.acra.2011.09.014.

Muttarak, M, P Kongmebhol, and N Sukhamwang (2009). "Breast calcifications: which are malignant". *Singapore Med J* 50.9, pp. 907–914.

Nguyen, Hieu T. et al. (2022). "VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography". *arXiv preprint*. DOI: 10.48550/arxiv.2203.11205.

Parikh, Jay R., Jia Sun, and Martha B. Mainiero (2020). "Prevalence of Burnout in Breast Imaging Radiologists". *Journal of Breast Imaging* 2.2, pp. 112–118. DOI: 10.1093/JBI/WBZ091.

Parmar, Chintan et al. (2018). "Data Analysis Strategies in Medical Imaging". *Clinical cancer research : an official journal of the American Association for Cancer Research* 24.15, pp. 3492–3499. DOI: 10.1158/1078-0432.CCR-18-0385.

Peintinger, Florentia (2019). "National Breast Screening Programs across Europe". *Breast Care* 14.6, pp. 354–358. DOI: 10.1159/000503715.

Qu, Wendi et al. (2020). "Assessing and mitigating the effects of class imbalance in machine learning with application to X-ray imaging". *International Journal of Computer Assisted Radiology and Surgery* 15.12, pp. 2041–2048. DOI: 10.1007/S11548-020-02260-6/FIGURES/3.

Rajpurkar, Pranav et al. (2022). "AI in health and medicine". *Nature Medicine 2022 28:1* 28.1, pp. 31–38. DOI: 10.1038/s41591-021-01614-0.

Rebolj, Matejka et al. (2018). "Addition of ultrasound to mammography in the case of dense breast tissue: systematic review and meta-analysis". *British Journal of Cancer* 118.12, pp. 1559–1570. DOI: 10.1038/s41416-018-0080-3.

Ribli, Dezso et al. (2018). "Detecting and classifying lesions in mammograms with Deep Learning". *Scientific Reports* 8.1, p. 4165. DOI: 10.1038/s41598-018-22437-z.

Rodriguez-Ruiz, Alejandro et al. (2019). "Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists". *Journal of the National Cancer Institute* 111.9, pp. 916–922. DOI: 10.1093/jnci/djy222.

Saito, Takaya and Marc Rehmsmeier (2015). "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets". *PLOS ONE* 10.3, e0118432. DOI: 10.1371/JOURNAL.PONE.0118432.

Schaffter, Thomas et al. (2020). "Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms". *JAMA network open* 3.3, e200265. DOI: 10.1001/jamanetworkopen.2020.0265.

Shen, Li et al. (2019). "Deep Learning to Improve Breast Cancer Detection on Screening Mammography". *Scientific Reports* 9.1. DOI: 10.1038/s41598-019-48995-4.

Shoeibi, Afshin et al. (2021). "Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: A review". *Computers in Biology and Medicine* 136, p. 104697. DOI: 10.1016/j.compbiomed.2021.104697.

Sisternes, Luis de et al. (2015). "A computational model to generate simulated three-dimensional breast masses". *Medical Physics* 42.2, pp. 1098–1118. DOI: 10.1118/1.4905232.

Stadnick, Benjamin et al. (2021). "Meta-repository of screening mammography classifiers". *arXiv preprint*. DOI: 10.48550/arxiv.2108.04800.

Sung, Hyuna et al. (2021). "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries". *CA: A Cancer Journal for Clinicians* 71.3, pp. 209–249. DOI: 10.3322/CAAC.21660.

Tardy, Mickael (2021). "Deep learning for computer-aided early diagnosis of breast cancer". PhD thesis. Nantes: Ecole Centrale de Nantes.

Tardy, Mickael and Diana Mateus (2021). "Looking for Abnormalities in Mammograms with Self- And Weakly Supervised Reconstruction". *IEEE Transactions on Medical Imaging* 40.10, pp. 2711–2722. DOI: 10.1109/TMI.2021.3050040.

— (2022). "Leveraging Multi-Task Learning to Cope With Poor and Missing Labels of Mammograms". *Frontiers in Radiology* 1, p. 19. DOI: 10.3389/fradi.2021.796078.

Tardy, Mickael, Bruno Scheffer, and Diana Mateus (2019). "Uncertainty Measurements for the Reliable Classification of Mammograms". *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 11769 LNCS. Springer, pp. 495–503. DOI: 10.1007/978-3-030-32226-7{_}55.

Taylor, Luke and Geoff Nitschke (2019). "Improving Deep Learning with Generic Data Augmentation". *Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence, SSCI 2018*, pp. 1542–1547. DOI: 10.1109/SSCI.2018.8628742.

Wing, Paul and Margaret H. Langelier (2009). "Workforce Shortages in Breast Imaging: Impact on Mammography Utilization". *American Journal of Roentgenology* 192.2, pp. 370–378. DOI: 10.2214/AJR.08.1665.

Wu, Eric et al. (2018). "Conditional infilling GANs for data augmentation in mammogram classification". *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11040 LNCS, pp. 98–106. DOI: 10.1007/978-3-030-00946-5{_}11/COVER/.

Wu, Nan et al. (2019). "Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening". *IEEE Transactions on Medical Imaging*, pp. 1–1. DOI: 10.1109/tmi.2019.2945514.

Zhou, S. Kevin et al. (2021). "A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises". *Proceedings of the IEEE* 109.5, pp. 820–838. DOI: 10.1109/JPROC.2021.3054390.

Zhou, Zongwei et al. (2019). "Models Genesis: Generic Autodidactic Models for 3D Medical Image Analysis". *Med Image Comput Comput Assist Interv.* 11767, pp. 384–393. DOI: 10.1007/978-3-030-32251-9{_}42.

Zhu, Wentao et al. (2017). "Deep multi-instance networks with sparse label assignment for whole mammogram classification". *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10435 LNCS, pp. 603–611. DOI: 10.1007/978-3-319-66179-7{_}69.