Masking in Mammograms: An Improved Risk Stratification Model using ResNet

A Thesis Submitted
in Partial Fulfillment of the Requirements
for the Degree of
Master of Technology
in
Software Engineering

by
Vysakh Ramesh
2020SW22

Under the Guidance of
Prof. M. M. Gore

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MOTILAL NEHRU NATIONAL INSTITUTE OF TECHNOLOGY ALLAHABAD
PRAYAGRAJ – 211004, INDIA
June, 2022

UNDERTAKING

I declare that the work presented in this thesis titled "*Masking in Mammograms: An Improved Risk Stratification Model using ResNet*", submitted to the Department of Computer Science and Engineering, Motilal Nehru National Institute of Technology Allahabad, for the award of the ***Master of Technology*** degree in ***Software Engineering***, is my original work. I have not plagiarized or submitted the same work for the award of any other degree. In case this undertaking is found incorrect, I accept that my degree may be unconditionally withdrawn.

June, 2022
Prayagraj

(Vysakh Ramesh)

CERTIFICATE

Certified that the work contained in the thesis titled "*Masking in Mammograms: An Improved Risk Stratification Model using ResNet*", by *Vysakh Ramesh*, Registration Number *2020SW22* has been carried out under my supervision and this work has not been submitted elsewhere for a degree.

(Prof. M. M. Gore)
Dept. of Computer Science and Engineering
MNNIT Allahabad

June, 2022

– ***Dedicated to all mothers***

Acknowledgment

My heartfelt gratitude to *Prof. M. M. Gore*, my thesis advisor, for being my life coach, pushing me to step outside of my comfort zone, teaching me how to become a professional, pointing out minor errors and raising our standards high, and for being the critic whose daily dose of motivation shaped the entirety of this thesis. Also, I extend my thanks to my fellow lab mates: *Uday Singh* for inviting me for lunches and giving me seldom breaks from the dismay of the hostel food, *Java Sonker* for her constant chit chats making the lab environment reverberant, the duo *Syed Farhat* and *Vikas Reddy* through their time at MNNIT, for being my go to person and sharing cock 'n bull stories respectively.

Getting through M.Tech and life in general at Allahabad required more than academic support, and I thank my best friend *Dimple Motwani* for listening to and, at times, having to tolerate me over the course of time. I cannot begin to express my gratitude and appreciation for her friendship. Words cannot describe my thanks to my partner *Abhirami Harish*, for her unwavering love, care, and support over the years, as well as her occasional reminders to stay on course with my goals. For taking time off her day to listen to my research problems, proofreading and feedback, she is the unsung heroine of my work, and I am sincerely looking forward to her endured support for my Ph.D.

Last but not the least, I would like to thank my family: my super mom *Sheeba Remesh* who sacrificed her career and took the greatest care of us three boys through our hardest times, my dad *Remesh Kakkanat* for being the father to his siblings and teaching us along the way, the values of hard work and unity. I thank my family for reminding me to take small strides and enjoy the life.

- Vysakh Ramesh

❝If we knew what it was we were doing, it would not be called research, would it?❞

- Albert Einstein

Abstract

Mammograms are an easy and cost-effective way to detect breast cancer early. Tumours in the mammographic screening of woman having high-density breast tissues are prone to get unnoticed. These dense tissues hide possible developing tumours, a phenomenon called masking which results in false negatives and misses early detection. Identifying mammographic masking early could help recommend suitable radiological screening methods. In contrast to the previous approaches of classifying mammograms based on breast density using relatively smaller dataset of labelled masking images, we aim to stratify the masking potential using Convolutional Neural Network and the CSAW-M dataset containing 10,200 images. Also, in this thesis, we document six mammographic datasets, including CSAW, CSAW-M, CBIS-DDSM, INbreast, MIAS and OMI-DB. We have experimented with two models, normal and ordinal classification model using pre-trained ResNet18 and ResNet34 architectures. Training on the new CSAW-M dataset, we observe that, our models identify masking on par with radiologists.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Preface

Breast cancer is the most widely diagnosed cancer in women globally, and also tops the list of cancer related deaths worldwide. It is estimated that breast cancer accounts for 2.1 million diagnosed cases and 627,000 deaths in 2018 globally. Although the incidence rate of breast cancer is lower in low HDI regions [2], the relative trend of breast cancer incidence is rising in India [3], see figure 1. Furthermore, in India, breast cancer is identified at a later stage [4]. Previous studies report that early detection has been linked to a better prognosis [5, 6]. Therefore, early detection and diagnosis are crucial for optimal patient care.

Mammography, ultrasound, MRI, thermal imaging, self breast examination are some methods for screening and diagnosing breast cancer. Despite their long clinical success of mammography, ultrasound and MRI, questions have been raised on the negative effects of radiation, increased false positive rates and cost respectively [7]. This resulted in the development of low-cost wearable devices to predict the size and location of the cancerous tissue using thermal tomography [8]. Such devices are a viable option for self breast examination in low-income countries. However, thermographic or thermal tomographic imaging in clinical practice should only be employed as an adjunct screening method and should not replace mammography for the early detection of breast cancer [9]. Thus mammography remains the de facto gold standard for screening breast cancer.

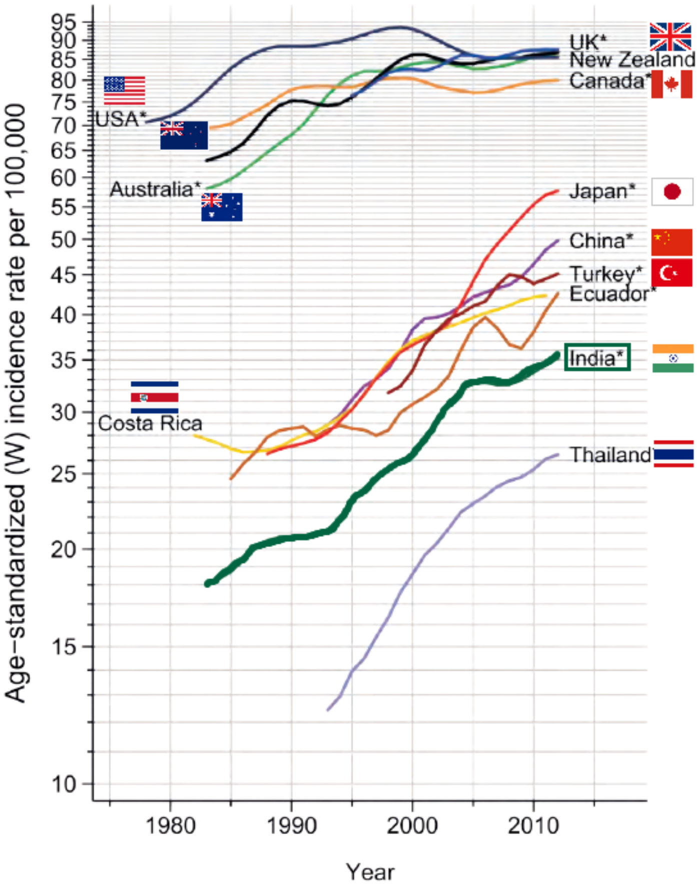


Figure 1: Beast Cancer Trends

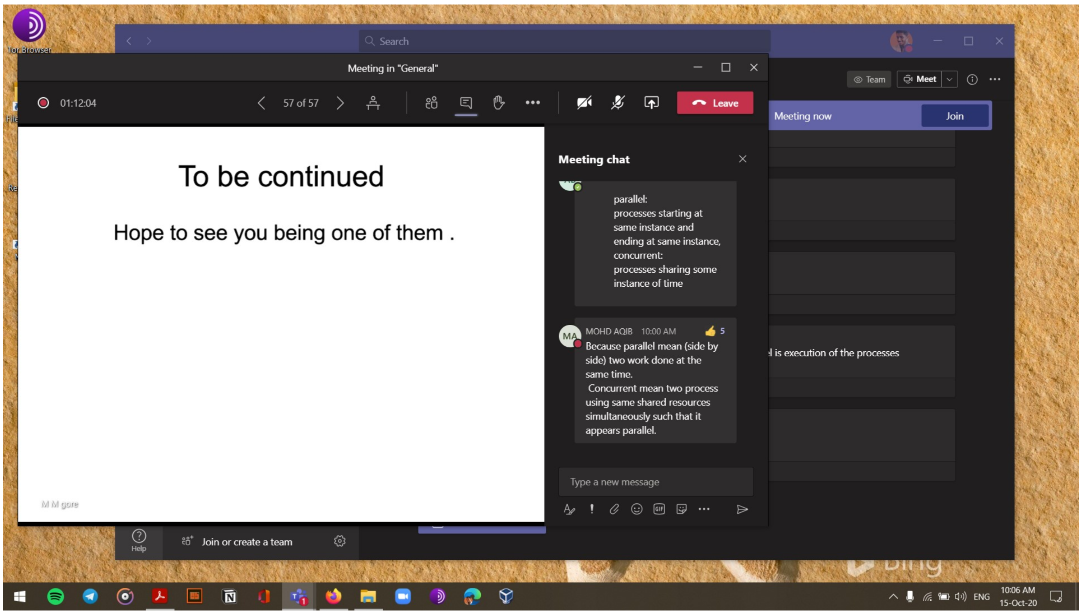
1.2 Masking

Tumours in high-density breasts are more likely to go unnoticed during mammographic screening. The dense tissues hide possible developing tumours, a phenomenon called masking which results in false negatives and misses early detection. Masking is further increased by fibroglandular tissues and overlapping tissues, reducing mammographic sensitivity. In fact, $7-22\%$ of women in the age group of $40-70$ have high density breasts in India [10]. The proportion is significantly higher in western countries. Hence, to recommend suitable screening method early identifying mammographic masking is necessary.

1.3 Motivation

I've nurtured the habit of compiling the curriculum vitae of my seniors, professors, and prominent researchers. Curiosity compelled me to investigate both their journeys and their publications. One thing I observed is that AI in Healthcare, Climate Change, and Ethics in AI are the most critical issues that they were addressing. Because the cause was worthwhile and the discipline was rather established, I was drawn to the domain at the confluence of AI and healthcare. Also, in the aftermath of the Covid-19 outbreak, healthcare research has become even more compelling. Then, I came across the paper *"CSAW-M: An Ordinal Classification Dataset for Benchmarking Mammographic Masking of Cancer"* that got accepted to the dataset track of NeurIPS 2020 conference. The availability of the dataset and 20 years of research in breast cancer masking gave me confidence to choose this topic.

Attending *Prof. M. M. Gore's* Research Methodology lecture and especially his class on Turing Award Winners was inspirational. The concluding slides of his lecture on Turing Award recipients on October 15, 2020, *"Hope to see you being one of them."* has motivated me to carry out this work.



To be continued

Hope to see you being one of them .

M M gore

Meeting chat

parallel:
processes starting at same instance and ending at same instance,
concurrent:
processes sharing some instance of time

MOHD AQIB 10:00 AM
Because parallel mean (side by side) two work done at the same time.
Concurrent mean two process using same shared resources simultaneously such that it appears parallel.

l is execution of the processes

Figure 2: CS21101 Research Methodology Class on Turing Award Winners

1.4 Objective

- Explore the publicly available mammogram datasets and summarise their features.
- Study the working of Convolutional Neural Network.
- Employ ResNet-18 and ResNet-34 architectures to identify masking potential.
- Compare the findings of our study with expert analysis.

1.5 Outline of Thesis

This report's remainder is organised as follows: The literature review is compiled in Chapter 2. Convolutional Neural Networks are discussed in Chapter 3 along with publicly available mammographic datasets for breast cancer detection and risk stratification. The proposed work is discussed in Chapter 4. The experimental setup is explained in Chapter 5, and the results are analysed in Chapter 6. The paper concludes with recommendations for future research in Chapter 7.

1.6 Summary

This chapter is an introduction to this report *"Masking in Mammograms: An Improved Risk Stratification Model using ResNet"*. The relevance of this topic is emphasised through the diagram that shows the rising breast cancer trend across the world. The chapter also briefly explains the phenomenon of masking in mammograms and the screening techniques used. The motivation behind this topic as well as the objectives are clearly mentioned in this chapter. Also, an outline of the report has been included.

Chapter 2

Preliminaries

2.1 Introduction

This chapter contains information that guides the reader to understand the preliminaries associated with this study. The history, architecture, and working principles of Convolutional Neural Networks are covered in this chapter. In addition, we have recorded 6 different publicly available mammographic datasets so that the reader can get a fundamental understanding of the properties of mammographic datasets.

2.2 Convolutional Neural Network

Convolutional Neural Network, also know as ConvNet or CNN, are a special kind of Artificial Neural Networks that can process data in grid like topology. Unlike the network's name, which implies the mathematical operation known as convolution it does not precisely correspond to the definition as in engineering or pure mathematics [11]. Convolution or convolving is the process of scaling down the input without loss of information. It is built on the ideas of parameter sharing, sparsity of connections and backpropagation.

The history of ConvNets goes back to the late 1980s when LeCun *et al.* [12] employed CNN to recognise handwritten zip codes sourced from the US Postal Service, thus creating LeNet. Due to the limitation of datasets and computational resources CNNs and the connectionism lost momentum until when Krizhevesky *et. al* [13] won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC-2012) [14] competition with his new architecture AlexNet. With ConvNets becoming more popular, many attempts have been made by the computer vision community to improve the performance of the original architecture resulting in ZFNet [15], GoogleNet [16], VGGNet [17], ResNet [18]. A concise summary of the different ConvNet architectures are shown in Table 1.

Name	Ref.	Year	Layers	Contribution
AlexNet	[13]	2012	8	Fast feature extraction, data compression, SVM classifier
ZFNet	[15]	2013	8	Deep feature visualization, feature invariance, feature evolution and feature importance
GoogleNet	[16]	2014	22	Global average pooling, auxilary classifier
VGGNet	[17]	2014	19	Using ReLU, increased network depth
ResNet	[18]	2015	118	Introduction of residual network,skip connections

Table 1: Popular CNN architectures [1]

The CNN architecture comprises of a series of convolutional, pooling and fully connected layers. The output of one layer is fed into the other sequentially. Apart from these three main components, the CNN also contain dropout layers and an output layer. These layers are briefly explained below:

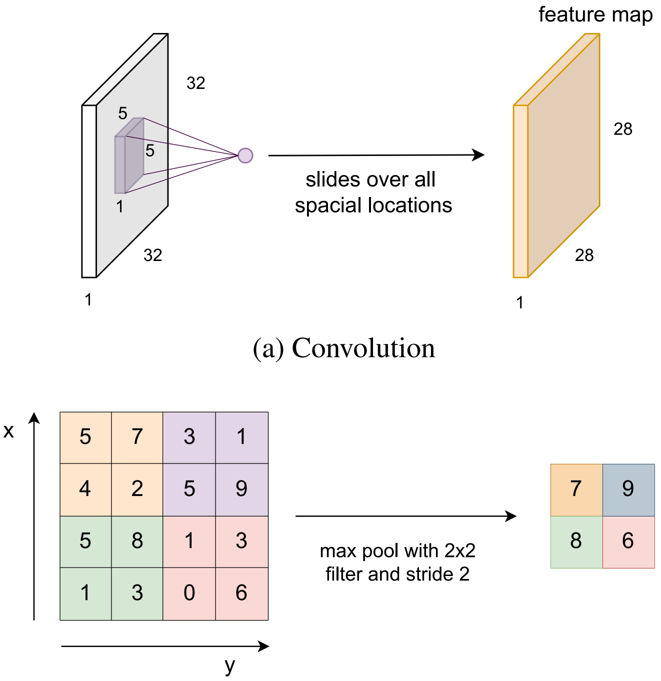
1. **Convolutional Layers**: The main task of the convolutional layer is feature extraction. CNN takes an $m \times n$ image as input and slides the kernel weights or convolution filter of dimension $f \times f$ over the input image. The convolution operation involves the dot product sum of the input *I* and the kernel *K*, where *i* and *j* are the indices.

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m,j-n)$$

The resultant 2D output is called feature map or feature vector. Initially the kernel weights are initialised at random, then modified later throughout the training process. Figure 3a illustrates the convolution process.

2. **Pooling Layer**: The responsibility of the pooling layer is to down-sample the input feature vector without the loss of information. The Pooling layers are often affixed between convolutional layers to reduce the overall number of parameters. The average, maximum, and minimum pooling methods are the most well-known and often used pooling algorithms. Figure 3b illustrates the max pooling operation.

3. **Fully Connected Layer**: The fully connected or FC-layer forms the last few layers of the ConvNet. Much like the conventional neural networks, nodes in an FC-layer have mappings to all the activations of the previous layer in a mesh-like topology. This structure makes it easy to use matrix multiplication and bias offset to determine the activations.



(a) Convolution

5	7	3	1
4	2	5	9
5	8	1	3
1	3	0	6

7	9
8	6

(b) Pooling

Figure 3: Operations in CNN Layers

Since 1996, CNNs have been employed in medical image processing. In his work Sahiner *et al.* [19] demonstrate the capability of CNN in classifying the ROI's on mammograms as either mass or normal tissue. Anthimopoulos *et al.* [20] apply ConvNets to classify interstitial lung cancer patterns. Setio *et al.* [21] won the Brain Tumour Segmentation Challennge 2013 using CNNs in MRI images. Setio detected pulmonary nodules from CT images using multi view CNNs. Recently, Ronneberger *et al.* [22] developed U-Net that employs ConvNets for biomedical image segmentation.

Although convolutional neural networks have achieved great success, there are some inherent limitations. To obtain good performance, ConvNets require a huge amount of training data. Being a deep learning algorithm, the training process is time consuming and requires heavy computational power. In addition, annotated medical imaging datasets are scarce unlike the natural image. Also, there exists class imbalance in the medical datasets which reflects the occurrence and trends of diseases. They do not have explicit internal representations of entities and their relationships. Furthermore, convolutional neural networks are incapable of comprehending images in terms of objects and their components. They recognise them as blobs of pixels with different patterns. They don't have internal representations of entities and their relationships that are explicit.

2.3 Dataset

Deep learning is notable for its ability to benefit from vast amounts of data. However, the lack of datasets in medical imaging is a major challenge for researchers. Quoting the statement on the challenges of procuring datasets from Greenspan *et al.*'s [23] work: "First, it is difficult to obtain funding for the construction of data sets. Second, scarce and expensive medical expertise is needed for high quality annotation of medical imaging data. Third, privacy issues make it more difficult to share medical data than natural images. Fourth, the breadth of applications in medical imaging requires that many different data sets need to be collected."

To solve some of these problems, workshops at medical imaging conferences are organising challenges to assist researchers in publishing quality datasets and to foster state of the art algorithms to set new benchmarks. In addition, NeurIPS has also recently introduced a dedicated datasets and benchmark track.

In the next six subsections, we explore the publicly available breast cancer datasets for our study. Each section briefly explains the affiliated research group publishing the dataset, the privacy concerns, ease of use and the accessibility constrains, if any. Table 2 summarises the commonly available datasets in breast cancer research.

Name	Ref.	Origin	Age Group	Population	Features
CSAW-M	[24]	Sweden	$40-74$	21,436	Detection mode Point-wise masking rank Density attribute
CSAW	[25]	Sweden	$40-74$	499,807	Tumour Size Detection mode Invasiveness Lymph node status Histological information Receptor status
OMI-DB	[26]	UK	$50-70$	173,319	Detection mode Invasiveness Pathological information Clinical information
MIAS	[27]	UK	NA	161	Information on background tissue Abnormalities Coordinates of centre of abnormality
CBIS-DDSM	[28]	USA	NA	118	Pathological Information Detection mode Improved ROI segmentation Rating of subtlety of abnormality Density attribute Patient age
INbreast	[29]	Portugal	NA	115	Abnormalities Clinical information Density attribute

Table 2: Mammography datasets

2.3.1 Cohort of Screen Aged Women

Cohort of screen aged women (CSAW) [25] is a collection of 2 million mammographic images of women aged between 40-74 years. The dataset is procured from screening at three breast cancer centers in Sweden . It contains 10,582 mammograms of women diagnosed with breast cancer. CSAW-S, a subset of the dataset contains information about cancer diagnosis, staging, and tumor characteristics as well as surgical characteristics, radiological assessments, and image acquisition metadata. Radiologists have pixel level annotated the tumours in mammograms and each diagnosis has been biopsy verified. The dataset has also been approved by ethical review board for AI research. Apart from being a benchmark to evaluate deep learning models, the potential applications of CSAW include creation of risk prediction network, tumour detection network and sensitivity assessment network. The dataset CSAW-S, can be obtained by submitting a request to the dataset sharing platform Zenodo with a valid justification. It is also being used by 4 other research groups.

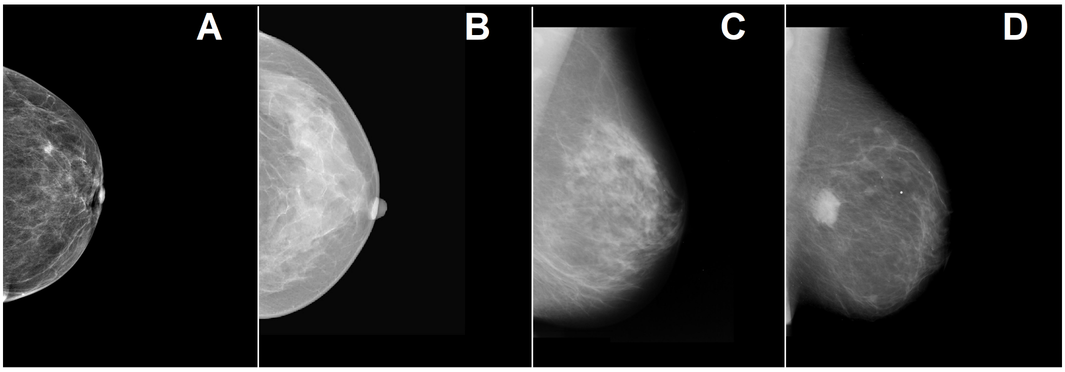


Figure 4: Mammography Images
(**A**) CSAW-M; (**B**) MIAS; (**C**) INbreast; (**D**) CSBIS-DDSM

2.3.2 Cohort of Screen Aged Women-M

Moein *et al.* [24] introduces CSAW-M, a subset of CSAW dataset containing 10,200 labelled mammographic masking images. These images are 1-2 orders larger in magnitude than the previous benchmark. With eight levels of graded masking potential examined by five radiological experts, CSAW-M could direct potential research in masking. In addition, the authors have experimented with building two baseline ResNet models to classify masking potential and identify large and invasive cancer. The dataset is developed by KTH Royal Institute of Technology in collaboration with Karolinska Institute and 2 other hospitals. Although the collected dataset is quite large, the dataset was not publicly available until December 8th. According to the authors, the dataset is hosted in SciLifeLab.

2.3.3 Curated Breast Imaging Subset-DDSM

Curated Breast Imaging Subset DDSM (CBIS-DDSM) [28] is a renewed rendition of the Digital Database for Screening Mammography (DDSM). It is a standard dataset to evaluate the performance of computer-aided detection and diagnostic systems. It has 753 cases of calcification and 891 cases of mass, whereas DDSM has a collection of 2,620 mammographic images. Techniques like image processing, image decompression, cropping, automated segmentation and removal of questionable images from the original DDSM dataset has been done to create the dataset. Other metadata include BI-RADS descriptors for mass shape, mass margin, calcification type, calcification distribution, breast density, rating of subtlety of abnormality and patient age. With updated ROI segmentation and reannotation by mammographers, the CBIS-DDSM is a standardized mammography dataset. The source code for creation of the dataset is publicly available.

2.3.4 INbreast

Moreira *et al.* [29] presents INbreast, a collection of 410 mammographic images procured from the Breast Centre in CHSJ Porto, Portugal. The annotations were made by an expert mammographer and validated by two others. Normal mammograms, mammograms with masses, mammograms with calcifications, architectural distortions, asymmetries, and images with multiple findings are all included in the database. The Portuguese National Committee of Data Protection and Hospital's Ethics Committee has approved the collection of mammograms. Although information regarding patient's age, family history, ACR breast density annotation and BI-RADS classification is provided in the dataset, all confidential medical information has been removed.

2.3.5 Mammographic Image Analysis Society

Mammographic Image Analysis Society (MIAS) [27] is a group of UK research groups dedicated to better the understanding of mammograms. They have created a digital mammogram database and is one of the earliest public datasets in mammography images. Films from the UK National Breast Screening Program were digitised to a pixel edge of 50 microns. There are 322 digitised films in the database. It also contains the truth-markings of the radiologist on the locations of any abnormalities that may be present. The database has been padded and reduced to a 200 micron pixel edge, resulting in images that are all 1024×1024. Mammographic images are available through the University of Essex's Pilot European Image Processing Archive (PEIPA).

2.3.6 OPTIMAM Mammography Image Database

The OPTIMAM Mammography Image Database (OMI-DB) is a collection of relational databases and a cloud storage system that houses mammographic images as well as clinical and pathological data [26]. It is made up of 2.5 million images collected from women aged 50 to 70 who were screened at three UK breast screening centres. The screenings were conducted as part of the National Health Service Breast Screening Program (NHSBSP) in the country. The database contains both processed images as well as unprocessed images with information regarding the findings of screening; women having normal breast, interval cancer, benignancy and malignancy. The data provided by OPTIMAM is used by over 30 research groups and companies [30]. OMI-DB has been accepted by the NHS Research Agency's ethics committee and is currently hosted in a website with proper documentation.

2.4 Summary

In this chapter we covered the history, architecture and working principles of convolutional neural networks. Also we have documented 6 different mammographic datasets that are publicly available.

Chapter 3

Literature Review

In the past two decades, a number of researches in the area of breast cancer has lead to an emphasis on risk stratification and tumour detection models. Risk stratification models calculate the likelihood of a person developing breast cancer. For this evaluation, the model requires details of the patient's age, race, ethnicity, heredity, diet, number of births, breast density and interval cancer data of those diagnosed with breast cancer. These models assist the high-risk patients in detecting breast cancer at an early stage. Tumor detection models, on the other hand, determine the probability of an individual having breast cancer during the screening process.

Wu *et al.* [31] have developed a two-stage CNN model that leverages both the global and local features of the mammographic images to classify breast cancer screening into malignant and benign. A redesigned ResNet architecture is used in both the pixel level and breast level models, which correspond to local and global information, respectively. They have demonstrated the feasibility of training the model with over one million high resolution mammograms. The authors credit the model's performance to the strategy of feeding the breast level model, a heatmap generated by the patch-level model. McKinney *et al.* [32] did similar work, where they constructed an AI system to detect breast cancer using clinical records from the United Kingdom and the United States. They validate their model with a comparison readers study which shows that the AI system exceeds the performance of radiologists. Training the model with the UK dataset and testing it on the US dataset, the authors demonstrate that their model generalises to unseen data.

Yala *et al.* [33] present a deep learning model trained on mammographic datasets that can estimate the risk of breast cancer. They have created three models: 1) risk factor based logistic regression model, 2) image only deep learning model, 3) hybrid deep learning model and demonstrates that their hybrid model outperforms the deep learning model in both white and coloured subgroups as well as a comparative breast density model. Hinton *et al.* [34], in their work, investigate and compare the performance of the deep learning network in classifying prior negative mammograms into interval and screen detected cancer with a classifier trained on BI-RADS density. To improve the performance of the model, they have employed the techniques of transfer learning, data preprocessing, data augmentation and hyperparameter sweep. They have made use of RestNet50 architecture.

BI-RADS is a common standard for recording abnormalities discovered on mammography, MRI, and ultrasound [35]. It allows for clear communication amongst all stakeholders involved. The writers of the fifth edition of BI-RADS, document the changes and compare it with its predecessor. The paper has three sections, each outlining the modifications made on mammography, MRI and ultrasound. The document summarizes the revisions so that they could be used as a reference for anyone using the dataset.

Boyd *et al.* [36], in their work conducted a case control study and examined the association of mammographic density with the risk of breast cancer. Their findings indicate that extensive mammographic density is related with an elevated risk of breast cancer, independent of whether the cancer was found by screening or other means. Boyd *et al.* [37] in an extension of the above work, further reviews the association of breast cancer with mammographic density. They discuss the biological causes of cancer risk, contrasting the risk factor of dense breasts with other risk factors like menstrual, reproductive, and familial risk. They also discover that, unlike most breast cancer risk factors, percentage mammographic density (PMD) can be altered. The paper also discusses the future prospects of PMD.

The review paper by Nazari *et al.* [38], also summarizes the current understating of the correlation between breast cancer and mammographic density. The authors point out that race, ethnicity, heredity, diet and number of births are some variables that influence the mammographic density. According to the study, women with dense breast tissue face two major issues: 1) An increased risk of breast cancer. 2) Late diagnosis of breast cancer due to mammographic screening's low sensitivity.

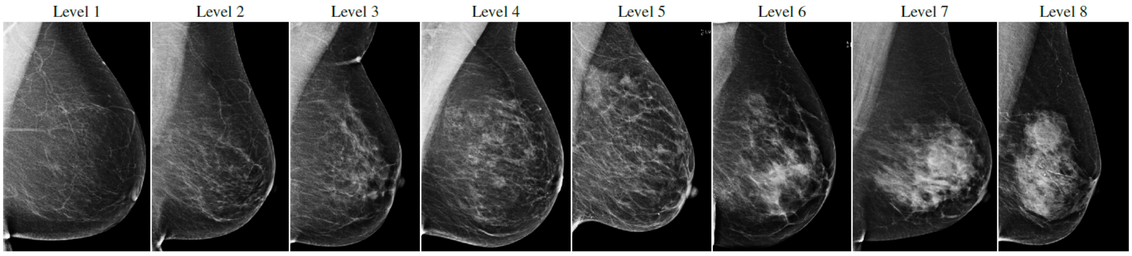
In their paper, Lehman *et al.* [39] develop and measure the acceptance of deep learning model to assess breast density in a real-time clinical environment. Their goal is to standardize mammographic density assessment, which is currently subjective and varies greatly amongst radiologists. The model classify the mammograms into the BI-RADS density groups. In addition, the model divides the images into two categories: dense and non-dense. They also ran a reader experiment to compare the results to the Deep Learning model. Maghsoudi *et al.* [40] present Deep-Libra, yet another method for estimating breast density from digital mammograms using artificial intelligence. Deep-Libra is a pipeline of AI modules that sequentially performs the following three computational steps: 1) segmentation for background removal, 2) segmentation for pectoralis muscle removal and 3) dense tissue segmentation and breast density calculation. The first two methods are based on deep learning and the last one is based on radiomic machine learning.

Ha *et al.* [41] demonstrates the use convolutional neural network to classify mammograms into high risk and low risk independent of the breast density. They employ a pixel-by-pixel heat map to highlight subregions in mammograms based on the most prevalent pattern observed in normal and high risk patients. In this retrospective case control study, the authors have highlighted the limitations of the dataset's size and procurement method. Cleland *et al.* [42] claims that radiologists can improve their assessment of mammographic masking with the help of models trained on Convolutional Neural Networks. The two models developed by the researchers at the Sunnybrook Research Institute to predict the risk factors are CNN-mask and CNN-mask2. Both use a pretrained VGG network to train the models on the masking dataset. Although the latter is trained on the BI-RADS metrics through an intermediate process, the results were not promising. The authors conclude that training on an improved and larger dataset can devise a better model.

In the paper [43], the authors investigate and explore different textural features that could improve the prediction of the risk associated with non-screen detected cancer along with breast density. The author asserts that the density metric in the BI-RADS dataset alone does not suffice to identify masking. The authors have addressed the limitation on the dataset's size and the density assessment rules employed. Holland *et al.* [44] have used volumetric breast density maps to quantify the masking risk in breast cancer.

In their work [45], the researchers at the Stanford ML group demonstrate that deep learning models learn better using weights pretrained on MoCo-CXR. To take advantage of the large unlabeled Chest X-Rays, they apply a technique called Momentum Contrast (MoCo). They show that the model outperforms the ImageNet pretrained counterpart in detecting the pathogens using this approach. Their findings show that self-supervised techniques can be extended to other medical imaging tasks.

Figure 5: Masking Potential



Eight different levels of masking potential cascaded from left to right. Level 1 being the least and level 8 being the highest. Image credit: Moein *et al.* [24]

Chapter 4

Estimation of Mammographic Masking

4.1 Overview

We employed two different classification approaches to estimate the levels of masking in mammograms, namely categorical classification and ordinal classification.

- Nominal Classification: In this method, we employ the standard categorical classification strategy, in which each class is treated separately.

- Ordinal Classification: The main goal of this method is to take into account the ordinal relationship between masking levels. This is a multi-label classification approach in which classifying an item into K ordinal classes is comparable to solving $K - 1$ independent binary classifications, with each class being a superset of the previous.

4.2 Architecture

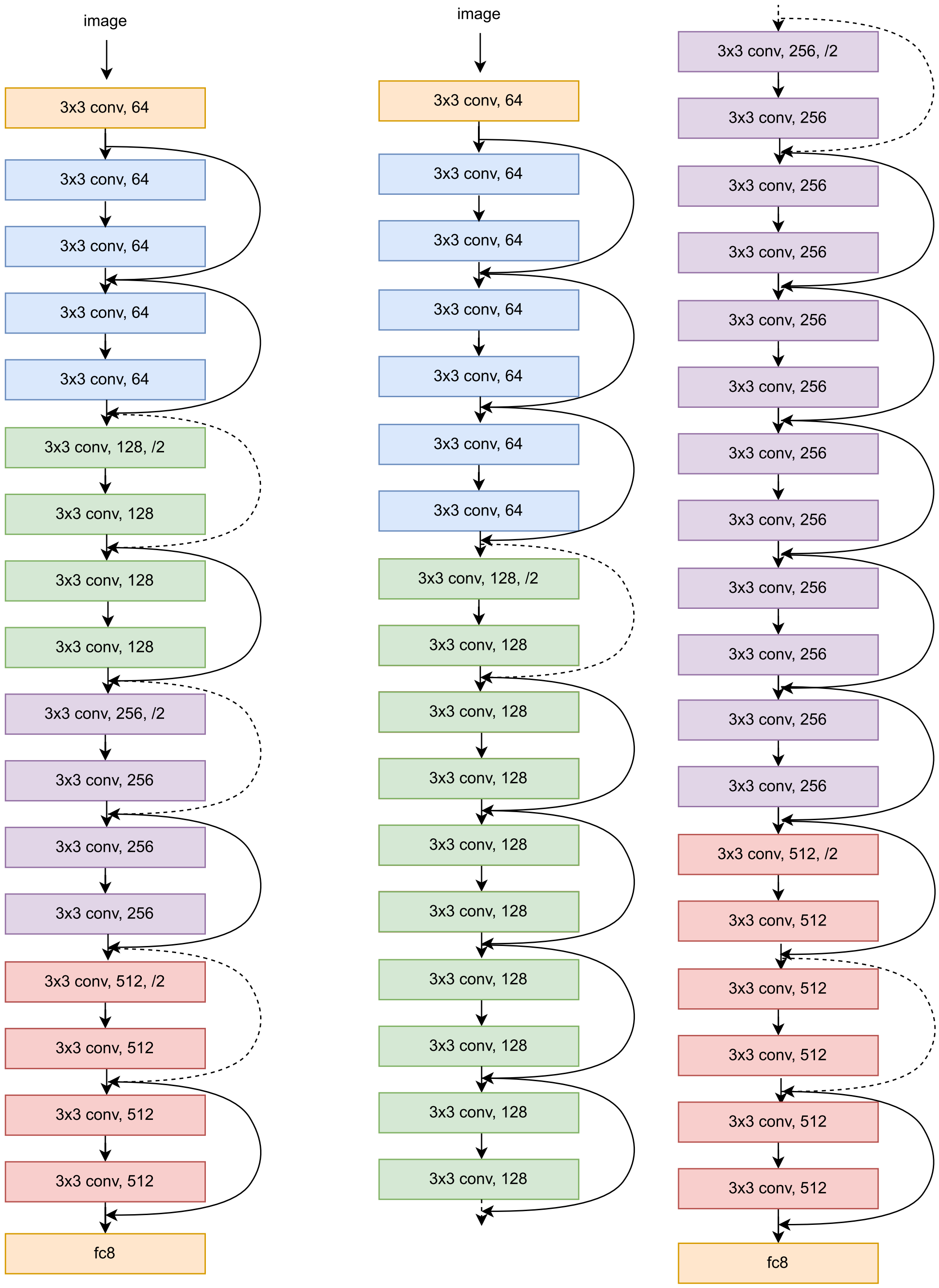
We have used two simple ResNet architectures namely ResNet-18 and ResNet-34 as backbone to create both the nominal classification and ordinal classification models. The architectures of ResNet-18 and ResNet-34 are depicted in the Figures 6a and 6b respectively. Apart from these, there are different variants of ResNets having 50, 101, 110, 152, 164, 1202 layers. We have chosen the ones having minimal number of layers considering the limitation of training time and computational resources. All the four models were pretrained with Imagenet dataset. Pretrained models have better learned features and have found to improve the performance.

Both the nominal and ordinal models used softmax layers at the end. The batch size was set to 64, Adam was used as the optimizer function with learning rate set to 1e-6. Categorical cross entropy loss was used in nominal classification, meanwhile binary cross entropy loss was used in ordinal classification. To further improve our results, we employed basic data augmentation techniques like random horizontal and vertical flips, rotations of $10°$ and $15°$, colour and contrast modifications on the already preprocessed dataset. Both the models were trained for 40 epochs. Each epoch consists of 569 steps, which are used to keep track of the checkpoints for evaluation. Checkpoints were made after every 50 steps.

The final masking rank is predicted using a method called masking score which calculates a cumulative score taking into account of probabilities of all 8 labels. For categorical model the score is calculated using the following formula:

$$\frac{1}{8}\sum_{l=1}^{8} p_l.l$$

where p_l represents the probability of the corresponding prediction and l for each of the masking levels 1-8 at the final softmax layer. In ordinal classification model to get continuous scores ,we first need to convert the cumulative probabilities that each output head produces to actual masking level probabilities p_l. Each output head models an individual binary classifier. For an input image x with true label L, each output head independently produces $P_k = p(L > k)$ where $k \in \{1, 2, 3, 4, 5, 6, 7\}$, denoting the probability that L exceeds k. Each output head is an independent binary classifier, and their probabilities may or may not be monotonic, thus we convert them to monotonic, i.e., for each $k > k$, it holds that P_k^k. At this stage, individual masking level probabilities is calculated by subtracting consecutive cumulative probabilities such that for each $k^{'} = k + 1$ it holds that $p_k = P_k P^{'}_k$. We could use the above formula to calculate the continuous score once we have analysed individual masking probability.



(a) ResNet-18

(b) ResNet-34

Figure 6: Redesigned ResNet Architecture

Chapter 5

Experimental Setup

5.1 Introduction

The first section of this chapter provides an overview of the CSAW-M dataset. Further sections describe the dataset features used, the train test split employed, and the ground truth chosen. Following that, the software modules and hardware specifications requirements, as well as the libraries utilised and their versions are detailed. The study was conducted in the form of a series of experiments, with data being sourced from the regular mammography screening between 2008 and 2015 at Karolinska University Hospital.

5.2 Dataset

CSAW-M was chosen from the six datasets listed in Section 2.3. The dataset consists of a total of 10,020 images split into 9,523 images of train data and 497 images of test data. Each set has a corresponding metadata file in csv format, refer Table 3 and Table 4. Of all these features only *Filename*, *Label*, *Expert_ID* were used from training set. Furthermore, the test set contains all of the expert values. The dataset stands out to be the only ordinal dataset having exclusive masking labels. The dataset can be used for: 1) masking models, 2) risk prediction models. Since the collection contains mammograms of breasts that do not have tumours, it cannot be used for cancer detection. The median was chosen as the ground truth label as it is 1) resilient to outliers, 2) always gets selected if there is a majority vote, and 3) simplifies the process of ranking masking levels. The dataset was sourced after agreeing to the terms and conditions and is stored in the google drive.

Filename	Image file name
Label	Annotation made by one of the experts, discrete number from 1 to 8
Expert_ID	Expert ID, from 1 to 5

Table 3: CSAW-M train metadata csv file

Filename	Image file name
Label	Ground-truth (median of 5 annotations), discrete number from 1 to 8
Expert_1	Annotation made by expert 1, discrete number from 1 to 8
Expert_2	Annotation made by expert 2, discrete number from 1 to 8
Expert_3	Annotation made by expert 3, discrete number from 1 to 8
Expert_4	Annotation made by expert 4, discrete number from 1 to 8
Expert_5	Annotation made by expert 5, discrete number from 1 to 8

Table 4: CSAW-M test metadata csv file

5.3 Software/Hardware

With an Nvidia Tesla V4 GPU, all the models were trained and tested on Google Colab for a total of 8 hours. For training, the deep learning framework PyTorch 1.11.0+cu113 was utilised. The findings were plotted using Matplotlib 3.2.2 and Seaobrn 0.11.2. Other significant frameworks used for the implementation were NumPy 1.21.6, Pandas 1.3.5, and Scikit-learn 1.0.2.

5.4 Summary

This chapter presents an overview of the dataset CSAW-M employed for the study, along with its intricacies. Further, the metadata of CSAW-M has been explained with the help of a table. The later section documents the software and hardware configurations required for the successful implementation of the model.

Chapter 6

Results and Analysis

6.1 Evaluation Metrics

The following metrics were used to assess the model's performance and compare it to expert predictions. These measures were chosen based on the authors' recommendations for the CSAW-M dataset [24].

- **Kendall's τ**: Kendall's τ is used to measure ordinal association between two variables. When samples have a comparable rank between the two variables, Kendall correlation will be high, and when samples have a different rank between the two variables, Kendall correlation will be low. Suppose two observations (x_i, y_i) and (x_j, y_j) are *concordant* if they are in the same order with respect to each variable. They are *discordant* if they are in the reverse ordering for X and Y or the values are arranged in opposite directions. The two observationsre tied if $x_i = x_j$ and/or $y_i = y_j$. The Kendall τ_B for measuring order association between variables X and Y is given by the following formula:

  $$\tau_B = \frac{P - Q}{\sqrt{(P + Q + X_0)(P + Q + Y_0)}}$$

  where P is the number of concordant pairs, Q is the number of discordant pairs, X_0 is the number of pairs tied only on the X variable, Y_0 is the number of pairs tied only on the Y variable. Kendall's τ ranges from -1 to $+1$, where -1 means perfect inverse correlation, 0 indicates no correlation and 1 indicates perfect correlation.

- **Average Mean Absolute Error (AMAE)**: AMAE is the average of the Mean Absolute Error (MAE) throughout the classes. MAE is defined as the average deviation of the predicted class from the ground truth. The MAE for the j^{th} class is calculated using the following formula:

  $$MAE_j = \frac{1}{N_j} \sum_{i=1}^{N_j} |pred(x_i) - act(x_i)|$$

  where N_j is the number of samples in class j, $pred$ is the predicted value and act is the ground truth for sample x_i. Let L be the total number of classes in the dataset then, AMAE can be defined by the following formula:

  $$AMAE = \frac{1}{L} \sum_{j=1}^{L} MAE_j$$

  As a result of the class imbalance and ordinal nature of the dataset, AMAE is better metric to use than accuracy.

- **F_1 score**: F_1 score is a metric which combines both the precision and recall of a model. It is defined as the harmonic mean of precision and recall. The score is calculated using the following formula:

  $$2 \times \frac{precision \times recall}{precision + recall}$$

  where *precision* is the fraction of true positive examples among the examples that the model classified as positive. *Recall* is the fraction of examples classified as positive, among the total number of positive examples. We use F_1 score of the highest masking levels (7-8) and lowest masking levels (1-2) to identify the model performance in both high masking and low masking levels respectively. The masking levels were chosen based on expert recommendation [24].

- **Confusion Matrix**: Confusion Matrix is an $N \times N$ error matrix to evaluate the performance of the classifier, where N is the number of classes. Each instance of the rows correspond to the ground truth and each instance of the columns corresponds to the model prediction. The matrix compares the ground truth to the machine learning model's predictions.

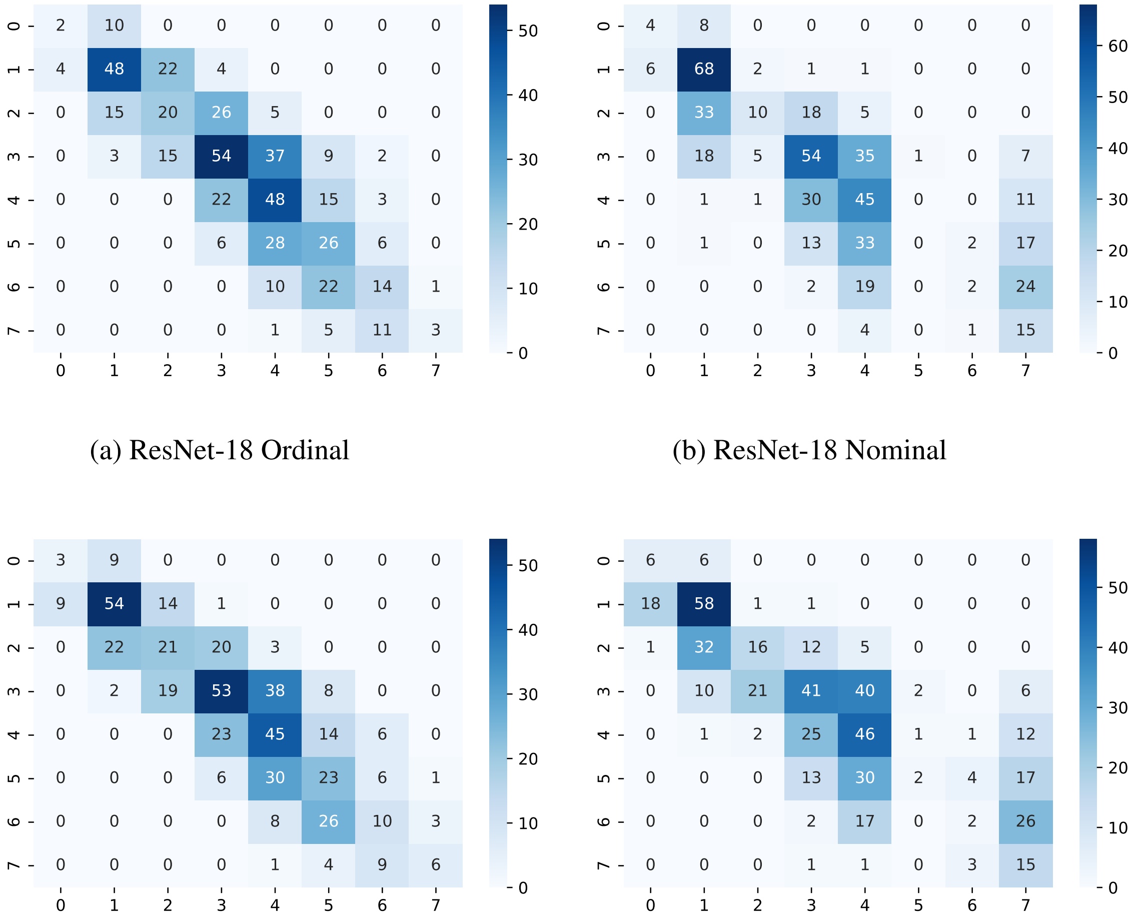
6.2 Results

The experts' agreement between ordinal classification of mammograms to the ground truth median as well as the models agreement to the ground truth is summarised in Table 5. Each columns represent the evaluation metrics and each row represents either the proposed model or the expert model. The top three values in each row is highlighted. As a rule of thumb, Kendall's $\tau_B \geq 0.3$ indicates a strong association.

From the Kendall's τ_B values in Table 3, it can be observed that both the ResNet models are having strong association scores, 0.7639 and 0.7444 as compared to the experts' whose highest score is 0.7232. Interestingly, ResNet models employing ordinal method has outperformed the expert analysis. Similarly, the AMAE observed for the Ordinal ResNet-34 model 0.7139, is lower than all other experts except Expert 1 for which the value is 0.6762. Also, it must be noted that the ResNet models using both the approach have AMAE < 0.88, whereas only 2 of the 5 experts' analysis stays in this range. In the metric F_1 score for low masking levels, both Nominal and Ordinal ResNet-34 having values 0.7928 and 0.7937 are very close to the highest expert score of 0.7940. ResNet-18 models have F_1 score of 0.7511 and 0.7442 which is higher than 60% of the expert analysis. Although F_1 score of the experts outperformed that of all the models models on high masking levels, their scores were comparable.

		Kendall's τb	AMAE	F_1 on level 1-2	F_1 on level 7-8
ResNet-	Nominal	0.6991	0.8763	0.7511	0.5753
18	Ordinal	**0.7444**	0.7532	0.7442	0.5421
ResNet-	Nominal	0.7159	0.8327	**0.7928**	0.6013
34	Ordinal	**0.7639**	**0.7139**	**0.7937**	0.5183
	Expert 1	**0.7232**	**0.6762**	**0.7940**	**0.6154**
	Expert 2	0.7279	**0.7167**	0.7465	**0.6316**
Experts	Expert 3	0.5450	1.0037	0.7363	0.5200
	Expert 4	0.5554	1.0390	0.5430	**0.6242**
	Expert 5	0.6342	1.0321	0.6885	0.5225

Table 5: Comparison of model performance with experts



(a) ResNet-18 Ordinal

(b) ResNet-18 Nominal

(c) ResNet-34 Ordinal

(d) ResNet-34 Nominal

Figure 7: Confusion Matrix

These results show ResNet-34 is better than ResNet-18 and ordinal classification is better than nominal classification for the prediction of masking potential. The confusion matrix for the both the Ordinal and Nominal ResNet models are depicted in Figure 7.

6.3 Summary

In summary, the results indicate that Ordinal ResNet-34 model can predict the masking potential in mammograms better than the experts in most instances.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

During mammographic screening, tumours in high-density breasts are more likely to go undiscovered. Identification of masking helps in recommendation of appropriate screening procedures and increases the likelihood of early cancer treatment if diagnosed. In conclusion, we reviewed the literature and discussed how Convolutional Neural Networks work. We examined previous research in the field and compiled a list of publicly available mammographic datasets. With ResNet-18 and ResNet-34 architectures, we constructed two models that can estimate the masking potential: nominal and ordinal classification models. We also compared our findings to expert analysis and found that in most cases, the Ordinal ResNet-34 model can predict the masking potential in mammograms better than the experts.

7.2 Future Work

Our future work will make use of the huge collection of other publicly available datasets together with contrastive learning approach. We also intend to apply segmentation techniques to help the radiologists identify masked areas and make quicker decision along with predicting the masking levels. We also plan to deploy a Generative Adversarial Network to address the issue of class imbalance.

References

[1] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, 29(9):2352–2449, 2017.

[2] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68, 11 2018.

[3] Meesha Chaturvedi, K. Vaitheeswaran, K. Satishkumar, Priyanka Das, S. Stephen, and A. Nandakumar. Time trends in breast cancer among indian women population: An analysis of population based cancer registry data. *Indian Journal of Surgical Oncology*, 6, 12 2015.

[4] Jasmine A. McDonald, Roshni Rao, Marley Gibbons, Rajiv Janardhanan, Surinder Jaswal, Ravi Mehrotra, Manoj Pandey, Venkatraman Radhakrishnan, Pooja Ramakant, Nandini Verma, and Mary Beth Terry. Symposium report: breast cancer in india—trends, environmental exposures and clinical implications. *Cancer Causes & Control*, 32, 6 2021. Late stage screening in India.

[5] Greta Carioli, Matteo Malvezzi, Teresa Rodriguez, Paola Bertuccio, Eva Negri, and Carlo La Vecchia. Trends and predictions to 2020 in breast cancer mortality in europe. *The Breast*, 36, 12 2017.

[6] Steven A. Narod, Javaid Iqbal, and Anthony B. Miller. Why have breast cancer mortality rates declined? *Journal of Cancer Policy*, 5, 9 2015.

[7] Andrew Karellas and Srinivasan Vedantham. Breast cancer imaging: A perspective for the next decade. *Medical Physics*, 35, 10 2008.

[8] Linta Antony, K Arathy, Nimmi Sudarsan, M N Muralidharan, and Seema Ansari. Breast tumor parameter estimation and interactive 3D thermal tomography using discrete thermal sensor data. *Biomedical Physics & Engineering Express*, 7, 1 2021.

[9] Ramesh Omranipour, Ali Kazemian, Sadaf Alipour, Masoume Najafi, Mansour Alidoosti, Mitra Navid, Afsaneh Alikhassi, Nasrin Ahmadinejad, Khojasteh Bagheri, and Shahrzad Izadi. Comparison of the accuracy of thermography and mammography in the detection of breast cancer. *Breast Care*, 11, 2016.

[10] Tulika Singh, Niranjan Khandelwal, Veenu Singla, Dileep Kumar, Madhu Gupta, Gurpreet Singh, and Amanjit Bal. Breast density in screening mammography in indian population - is it different from western population? *The Breast Journal*, 24, 5 2018.

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[12] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1, 12 1989.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60, 5 2017.

[14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[15] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing.

[16] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[19] B. Sahiner, Heang-Ping Chan, N. Petrick, Datong Wei, M.A. Helvie, D.D. Adler, and M.M. Goodsitt. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. *IEEE Transactions on Medical Imaging*, 15(5):598–610, 1996.

[20] Marios Anthimopoulos, Stergios Christodoulidis, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE Transactions on Medical Imaging*, 35(5):1207–1216, 2016.

[21] Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J. van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clara I. Sánchez, and Bram van Ginneken. Pulmonary nodule detection in ct images: False positive reduction using multi-view convolutional networks. *IEEE Transactions on Medical Imaging*, 35(5):1160–1169, 2016.

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[23] Hayit Greenspan, Bram Van Ginneken, and Ronald M. Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35:1153–1159, 5 2016.

[24] Moein Sorkhei, Yue Liu, Hossein Azizpour, Edward Azavedo, Karolinska Institutet, Karin Dembrower, Dimitra Ntoula, Anthanasios Zouzos, Fredrik Strand, and Kevin Smith. Csaw-m: An ordinal classification dataset for benchmarking mammographic masking of cancer. 2021.

[25] Karin Dembrower, Peter Lindholm, and Fredrik Strand. A multi-million mammography image dataset and population-based screening cohort for the training and evaluation of deep neural networks—the cohort of screen-aged women (csaw). *Journal of Digital Imaging*, 33, 4 2020.

[26] Mark D. Halling-Brown, Lucy M. Warren, Dominic Ward, Emma Lewis, Alistair Mackenzie, Matthew G. Wallis, Louise S. Wilkinson, Rosalind M. Given-Wilson, Rita McAvinchey, and Kenneth C. Young. Optimam mammography image database: A large-scale resource of mammography images and clinical data. *Radiology: Artificial Intelligence*, 3, 1 2021.

[27] P Suckling J. The mammographic image analysis society digital mammogram database. *Elsevier Sc. B. V.*, pages 375–386, 1994.

[28] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L. Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data*, 4, 12 2017.

[29] Inês C. Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria João Cardoso, and Jaime S. Cardoso. Inbreast. *Academic Radiology*, 19, 2 2012.

[30] Lucy M. Warren, Mark D. Halling-Brown, Louise S. Wilkinson, Rosalind M. Given-Wilson, Rita McAvinchey, Matthew G. Wallis, David R. Dance, and Kenneth C. Young. Changes in breast density. In Robert M. Nishikawa and Frank W. Samuelson, editors, *Medical Imaging 2019: Image Perception, Observer Performance, and*

Technology Assessment, San Diego, California, United States, 16-21 February 2019. SPIE, 3 2019.

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

[32] McKinney, Scott Mayer, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, and Trevor Back. International evaluation of an ai system for breast cancer screening. *Nature*, 577:89–94, 1 2020.

[33] Adam Yala, Constance Lehman, Tal Schuster, Tally Portnoi, and Regina Barzilay. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology*, 292:60–66, 2019.

[34] Benjamin Hinton, Lin Ma, Amir Pasha Mahmoudzadeh, Serghei Malkov, Bo Fan, Heather Greenwood, Bonnie Joe, Vivian Lee, Karla Kerlikowske, and John Shepherd. Deep learning networks find unique mammographic differences in previous negative mammograms between interval and screen-detected cancers: A case-case study. *Cancer Imaging*, 19, 6 2019.

[35] D.A. Spak, J.S. Plaxco, L. Santiago, M.J. Dryden, and B.E. Dogan. Bi-rads® fifth edition: A summary of changes. *Diagnostic and Interventional Imaging*, 98(3):179–190, 2017.

[36] Norman F Boyd, Helen Guo, Lisa J Martin, Limei Sun, Jennifer Stone, Eve Fishell, Roberta A Jong, Greg Hislop, Anna Chiarelli, Salomon Minkin, and Martin J Yaffe. Mammographic density and the risk and detection of breast cancer. *The New England Journal of Medicine*, 356:227–236, 2007.

[37] Norman F Boyd, Lisa J Martin, Martin J Yaffe, and Salomon Minkin. Mammographic density and breast cancer risk: current understanding and future prospects. *Breast Cancer Research*, 13, 12 2011.

[38] Shayan Shaghayeq Nazari and Pinku Mukherjee. An overview of mammographic density and its association with breast cancer. *Breast Cancer*, 25:259–267, 5 2018.

[39] Lehman, Constance D., Adam Yala, Tal Schuster, Brian Dontchos, Manisha Bahl, Kyle Swanson, and Regina Barzilay. Mammographic breast density assessment using deep learning: Clinical implementation. *Radiology*, 290:52–58, 1 2019.

[40] Omid Haji Maghsoudi, Aimilia Gastounioti, Christopher Scott, Lauren Pantalone, Fang-Fang Wu, Eric A. Cohen, Stacey Winham, Emily F. Conant, Celine Vachon, and Despina Kontos. Deep-libra: An artificial-intelligence method for robust quantification of breast density with independent validation in breast cancer risk assessment. *Medical Image Analysis*, 73:102138, 2021.

[41] Richard Ha, Peter Chang, Jenika Karcich, Simukayi Mutasa, Eduardo Pascual Van Sant, Michael Z. Liu, and Sachin Jambawalikar. Convolutional neural network based breast cancer risk stratification using a mammographic dataset. *Academic Radiology*, 26:544–549, 4 2019.

[42] Theo Cleland, James G. Mainprize, Anne L. Martel, Olivier Alonzo-Proulx, Martin J. Yaffe, Roberta A. Jong, and Jennifer A. Harvey. Use of convolutional neural networks to predict risk of masking by mammographic density. In *Medical Imaging 2019: Computer-Aided Diagnosis, San Diego, California, United States, 16-21 February 2019*, volume 10950 of *SPIE Proceedings*, page 69. SPIE, 3 2019.

[43] James Mainprize, Olivier Alonzo-Proulx, Taghreed I. Alshafeiy, James T. Patrie, Jennifer A. Harvey, and Martin J. Yaffe. Prediction of cancer masking in screening mammography using density and textural features. *Academic Radiology*, 26:608–619, 5 2019.

[44] Katharina Holland, Carla H. Van Gils, Ritse M. Mann, and Nico Karssemeijer.

Quantification of masking risk in screening mammography with volumetric breast density maps. *Breast Cancer Research and Treatment*, 162, 4 2017.

[45] Hari Sowrirajan, Jingbo Yang, Andrew Y. Ng, and Pranav Rajpurkar. Moco-cxr: Moco pretraining improves representation and transferability of chest x-ray models. 10 2020.