caption

university of groningen

faculty of science and engineering

University of Groningen

Breast Density Estimation in T1-weighted MRI using Deep Learning

Master's Thesis

To fulfil the requirements for the degree of
Master of Science in Artificial Intelligence
at the University of Groningen, under the supervision of
Prof. Dr Peter van Ooijen (Radiation Oncology, University Medical Center Groningen),
Maruf Dhali (Lecturer, Artificial Intelligence, University of Groningen), and
Xueping Jing (PhD candidate, University Medical Center Groningen)

Muskan Muskan (s4336496)

March 31, 2023

Contents

Acknowledgments

First and foremost, I am grateful to my external supervisor, Prof. Dr Peter van Oijenfor, for providing me with this opportunity to study breast MRI images at Data Science Center in Health (DASH), University Medical Center Groningen (UMCG). I am thankful to my daily supervisor, Xueping Jing, for his valuable feedback and insights throughout the research project. His understanding of the breast MRI images and deep learning were a great help in making me understand the project requirements and make quick progress in the initial phase of the project. I also sincerely thank my internal supervisor, Maruf Dhali, for his persistent guidance and feedback throughout the study. His enthusiasm and quick availability made all the administrative tasks run smoothly. I am thankful to my colleague at DASH, Luis, for lending me his computer part-time to perform my initial experiments. A special thanks to my academic advisor, Dr Rachel van der Kaaij, for being invariably approachable and helpful. I would also like to thank my dear friends, Siemone Zuidema and Jorn Flach for their ceaseless support and understanding. Many thanks to Eshita Sengupta, and Varun Ravi Varma for brainstorming ideas with me. I am eternally grateful to my parents, Ramesh Kumar and Shilpi Sangwan for always providing for me, trusting in my capabilities, and making my dream of studying abroad a reality. A special thanks to my brother, Madhur, for being with me during these challenging times.

Abstract

The ratio of fibroglandular tissue to fatty tissue determines breast density. It is divided into four categories, almost entirely fatty, scattered fibroglandular tissue, heterogeneously dense, and extremely dense. Higher breast density increases the risk of breast cancer. It is one of the most commonly diagnosed cancers in women worldwide. The screening for breast cancer is mainly done using mammography. However, the sensitivity of mammograms decreases in cases of high breast density. Further, mammograms can have high variability depending upon the body position while screening, compression levels, and intensity of x-rays. On the other, magnetic resonance imaging (MRI) is a 3-dimensional screening, hence, it can give a more consistent and accurate image of the breast. Additionally, MRI gives a high contrast of soft tissues between fibroglandular and fatty tissues. Hence, it is a better imaging technique for breast density estimation. Conventionally, breast density is estimated using visual assessment of the radiologists. However, for faster and more consistent results, deep learning methods have been studied. Most of these studies used mammograms, and only a few applied deep learning to MRI images. This study aims to perform breast density classification using deep convolutional neural networks (CNNs) on breast MRI. Additionally, it aims to learn the impact of different architectural choices such as using 3-dimensional vs 2-dimensional CNNs, giving higher weights to less represented classes (such as extremely dense), and the impact of adding transfer learning on the classification performance. A total of 960 breast MRI images were evaluated from 508 patients with a mean age of 45 years ± 11 (standard deviation). The input to the deep learning pipeline was T1-weighted sequences from MRI images. Various ResNet-18 architectures were experimented with, such as 3D ResNet-18 with 3 slices, or with the whole volume, 2D ResNet-18 with pre-trained ImageNet weights, and class weights. The 2D CNN with transfer learning and class weights outperformed all ($Accuracy = 0.83$, $AUC = 0.87 \pm 0.05$). It was closely followed by 2D CNN with transfer learning ($Accuracy = 0.78$, $AUC = 0.88 \pm 0.04$). These were followed by the 3D CNN with three slices ($Accuracy = 0.75$, $AUC = 0.85 \pm 0.05$), which outperformed the 3D CNN with the whole volume ($Accuracy = 0.70$, $AUC = 0.67 \pm 0.04$). Contrary to our assumption, the inclusion of all slices did not improve the performance of the model. It is suspected to be an effect of overfitting, as the complexity of the models was high (millions of parameters) and the training samples were scarce. The 3D CNN was most affected as it had 3x more parameters than the 2D CNN. As per the assumption, using transfer learning increased the performance of the 2D CNN. Lastly, it is inferred that deep learning can be used for automatic breast density classification with a reasonable agreement between predictions from CNN and the assessment of radiologists.

List of abbreviations

Abbreviation	Full form
NHS	National Health Service
ACS	American Cancer Society
WHO	World Health Organization
FDA	U.S. Food and Drug Administration
IARC	International Agency for Research on Cancer
EUSOBI	European Society of Breast Imaging
UMCG	University Medical Center Groningen
DASH	Data Science Center in Health
ACR	American College of Radiology
BI-RADS	Breast Imaging Reporting and Data System
DICOM	Digital Imaging and Communications in Medicine
NIfTI	Neuroimaging Informatics Technology Initiative
DM	Digital mammography
CT	Computed tomography
DBT	Digital breast tomosynthesis
MRI	Magnetic resonance imaging
FGT	Fibroglandular tissue
TR	Repetition time
TE	Echo time
DL	Deep learning
ML	Machine learning
AI	Artificial intelligence
GAN	Generative adversarial network
CNN	Convolutional neural network
ANN	Artificial neural network
ResNet	Residual network
FCM	Fuzzy c-means
EM	Expectation maximisation
DSC	Dice similarity coefficient
SMOTE	Synthetic Minority Oversampling Technique
ROC	Receiver operating characteristic
AUC	Area under the curve
ReLU	Rectified linear unit
CI	Confidence interval
TP	True positive
FP	False positive

TN	True negative
FN	False negative
TPF (or TPR)	True positive fraction (or True positive rate)
FPF (or FPR)	False positive fraction (or False positive rate)
TNF (or TNR)	True negative fraction (or True negative rate)
FNF (or FNR)	False negative fraction (or False negative rate)
PPV	Positive predictive value
NPV	Negative predicted value
2D CNN	2-dimensional convolutional-neural-network
2Dw CNN	2-dimensional class-weighted convolutional-neural-network
3D3 CNN	3-dimensional with 3 slices convolutional-neural-network
3D CNN	3-dimensional convolutional-neural-network

Table 1: The list of all the abbreviations used in the report along with their meanings.

1 Introduction

In radiography, the female breast shows two contrasting components namely, the fibroglandular tissue and the fatty tissue. The latter comprises fibrous connective tissue and glandular epithelial cells, present in the duct linings of the breast [1]. The measure of the amount of fibroglandular tissue present in the breast relative to the fatty tissue is known as breast density. It has been observed that breast cancer predominantly develops among the fibroglandular tissue. Consequently, breast density is an important risk factor for developing breast cancer. Further, it is independent of other risk factors such as family history [2]. Breast cancer is noted to be one of the most regularly diagnosed cancers in females. Further, it leads to 25.2% of all female deaths due to cancer, as per American Cancer Society [3] [4] [5]. In the USA, it is the second most commonly occurring cancer in women after skin cancer [3]. Further, since 2005 the country has seen an increasing incidence rate of breast cancer [6]. Whereas in the UK it is the most frequent cancer type at the moment [4]. Nonetheless, the Netherlands had the highest age-standardised incidence rate of breast cancer in the year 2006. Recently, in 2020, the country came second to Belgium [7]. Hence, further research in the field of breast cancer and its early detection is extremely important worldwide and especially in countries with high incidence and death rates.

A study published in 2002 has reported that women with a percentage of breast density more than 50% on a mammogram have a 1.8 to 6.0 times elevated risk of breast cancer development compared to other women of the same age but with lower breast density [8] [9]. In addition to breast density being a risk factor for the development of cancer, it may also mask some cancers during screening. This is due to the fact that both breast cancer and fibroglandular tissue appear white on a mammogram [9]. Mammography is used as a primary screening method for breasts, which results in 2- or 3-dimensional mammograms. However, as previously mentioned, the sensitivity of mammograms is low on dense breasts. Hence, in such cases, additional screening is advised. A breast MRI is often used for this purpose, as it performs relatively well even in the case of dense breasts. Recently, the European Society of Breast Imaging (EUSOBI) has also recommended breast MRI as an additional screening for women with extremely dense breasts [2]. Breast density is divided into four categories according to the Breast Imaging Reporting and Data System (BI-RADS) [10]. These are **a.** almost entirely fatty, **b.** scattered areas of fibroglandular density, **c.** heterogeneously dense and **d.** extremely dense, as shown in Figure 1.

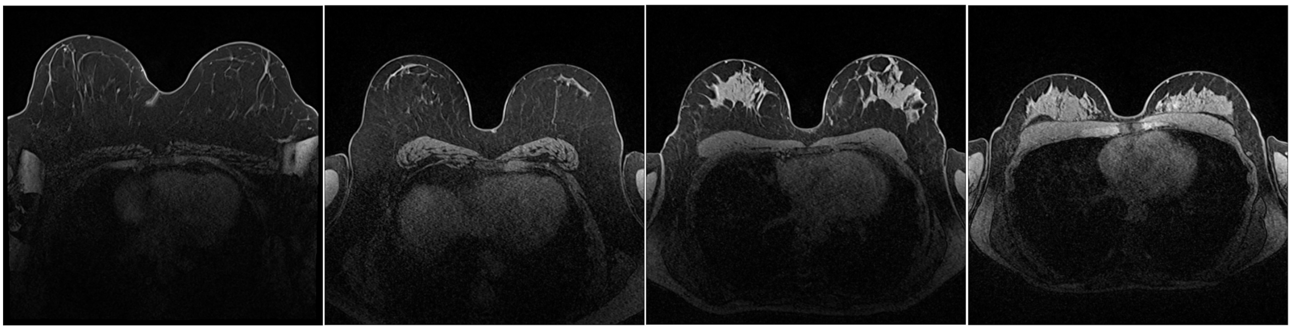


Figure 1: Examples of different breast compositions categorised into the following BI-RADS groups (a) almost entirely fatty; (b) scattered areas of fibroglandular density; (c) heterogeneously dense; (d) extremely dense [left to right].

Deep learning using convolutional neural networks is very popular with image classification tasks due to its high accuracy. Recently, it is also being widely used in the medical field to analyse diagnostic images. Its performance on these images has oftentimes outperformed the conventional methods of visual assessment. However, there are some challenges to using deep learning, such as requiring a large labelled dataset to train the convolutional neural networks. The collection of such a large dataset is often a big challenge in the medical domain. Mainly due to the extra security and separate ownership of data to different hospitals, as the data contains sensitive and personal information about the patients. To overcome the scarcity of data various methods are suggested such as transfer learning and data mining. In transfer learning, a network is used which has been pretrained on a different image dataset to improve and accelerate the feature learning process in the current image dataset. On the other hand, in the data mining approach, a publicly available dataset is automatically gathered to create reference labels for supervised training of the model [11].

Several methods have been suggested for breast density estimation and classification, such as computer-based quantitative methods [12], machine learning [13] [9], and deep learning [14] [15]. Deep learning has been broadly used for automatic image analysis in the medical domain. Many researchers have used it for breast density estimation using mammography [16] [17] [18] [11]. However, only some of them have used it on breast MRI images [15] [14]. Out of these, only a few estimated breast density without segmentation such as [14], where they used regression. To our best knowledge, there is no study that has done breast classification on MRI data without regression and segmentation. Hence, our study aims to perform breast density categorisation using deep learning on breast MRI without segmentation or regression. It also aims to compare different architectural choices, such as using a 2-dimensional versus 3-dimensional convolutional neural network. Further, the significance of using transfer learning is also studied.

1.1 Research questions

This study focuses on three research questions, which are described in detail in the following points.

- *Q1. How does a deep convolutional neural network (CNN) perform on breast MRI data?*

  The aim is to observe the performance of a standard deep learning model such as ResNet-18 convolutional neural network (CNN) [19]. on the breast magnetic resonance imaging data. First, prepare the dataset for training using image preprocessing techniques, such as normalisation, and resampling (downscaling). Moreover, enhance the performance using image augmentation using further image processing such as flipping, cropping, scaling, adjusting linear contrast and brightness, adding Gaussian noise, blurring with a Gaussian filter, and applying affine transformations. Further, enhance the performance using class weights by giving higher priority or weights to the less represented classes during the training of the model.

- *Q2. What are the implications of architectural choices, such as 2- versus 3-dimensional CNN?*

  Another aim of our study is to analyse the effect of using a 2-dimensional versus a 3-dimensional architecture of the ResNet-18 model. Further, compare the performance using standard metrics such as accuracy, precision, recall, and f1 score. Moreover, visually examine the performances using confusion matrices, and receiver operating characteristic curves.

- *Q3. Does transfer learning and the use of class weights improve the performance of CNN?*

  Our study further aims to handle the class imbalance issue and scarcity of data. More specifically, it aims to quantify the improvement in performance by using pretrained ResNet-18 using Imagenet dataset [20] and giving more priority to less-represented classes.

1.2 Thesis outline

The thesis is structured as follows. The first and current chapter i.e. *Introduction* briefly described breast density and its importance in relation to breast cancer. It also gave a clear description of the research questions. The second chapter i.e. *Clinical background* gives detailed statistics of breast cancer, its relation to breast density, and various screening methods. The third chapter i.e. *Related works*, will give an insight into the state of the art by analysing the background literature. In the fourth chapter i.e. *Methodology*, a detailed description of the methodology, starting with an analysis of the breast MRI data, some preprocessing techniques, and finally theory behind image classification using artificial neural networks. The fifth chapter *Experiment setup* gives details about our deep learning pipeline, and the technologies and resources used, further the configuration of MRI acquisition parameters and hyperparameters for the deep learning models. The sixth chapter *Results* showcases the results of our experiments, along with an explanation of how the results were obtained. In the seventh chapter, *Discussion*, the results from all the experiments are analysed. The eighth and final chapter *Conclusions* gives a summary of the entire study along with some drawbacks in our approach with subsequent suggestions for future research in the same direction.

2 Clinical background

The science behind breast density and its relation to breast cancer includes various terms and concepts. Some of these theories and terminologies are explained in this section. First, a brief statistical view of breast cancer is provided including the incidence and mortality rates in recent years. Second, the relationship between breast density and breast cancer is made clear with scientific reasoning. Third, various breast imaging techniques are illustrated, along with our motivation to use one of them.

2.1 Breast cancer statistics

According to a 2015 international study, breast cancer was found to be the most frequently occurring cancer and a major cause of cancer-related death in women worldwide [7]. This incidence rate was observed in 140 out of 184 countries globally and represented a quarter of all diagnosed cancers in women [21]. Another 2019 international study suggested the same that breast-cancer-related mortality has increased considerably worldwide [22]. This overall trend was observed in 195 countries from 1990 to 2015. The situation was worse in developing countries and low-income areas. The data used in the study was obtained from the Institute for Health Metrics and Evaluation (IHME) [23].

In the United States of America, breast cancer is the second most frequently found cancer after skin cancer in women, according to the American Cancer Society [3]. That is approximately 1 in 3 (30%) of all new cancers each year in women. The mortality rate of breast cancer is only second to lung cancer. Figure 2 shows the age-standardised rate of occurrence of new cases and mortality rates of breast cancer in the USA from 1975 to 2020. It is observed that each year from 2010 to 2019, the rate of new cases rose by 0.4% on average. However, the death rates fell by 1.3% each year between 2011 to 2020.

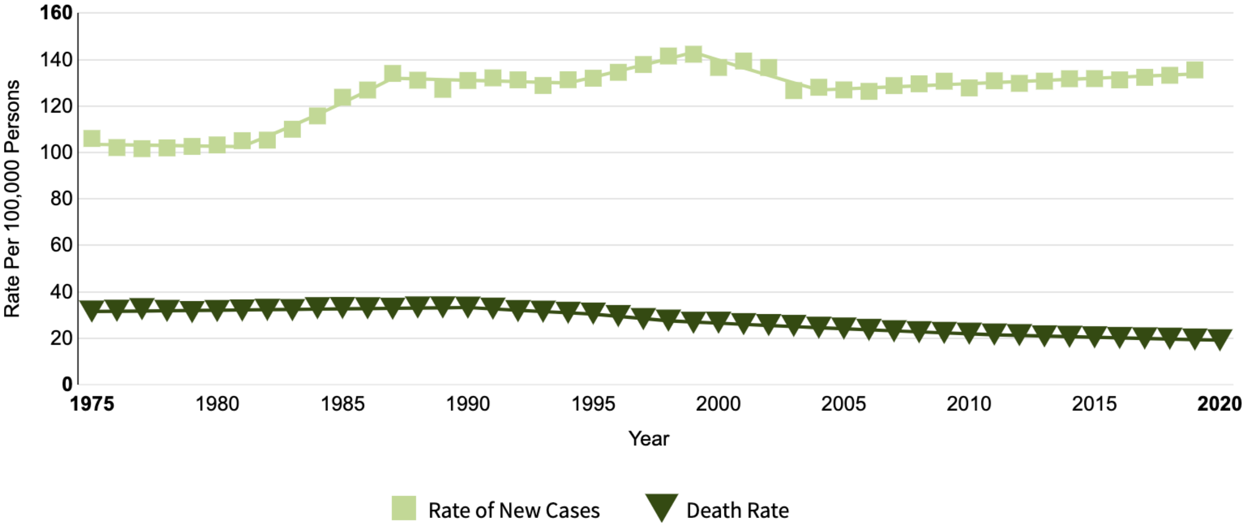


Figure 2: Trends in breast cancer in American women (source: National Cancer Institute, USA [6]).

Furthermore, in the United Kingdom, breast cancer is the most common type of cancer [4]. While according to the 2015 international study, **the Netherlands** ranked highest among the incidence of breast cancer from 2006 to 2007, with 95.3 cases per 100,000. However, it is the seventh on the morality rate list with 18.6 deaths per 100,000 [7]. However, recently in 2020, the Netherlands was second to Belgium in the incidence of breast cancer, with an age-adjusted rate of 100.9 per 100,000. This and other top ten incidence rates are shown in Table 2, along with their associated regions. Note, in the table, ASR denotes the age-standardised rates. The age-standardised rates represent a weighted average of the age-specific rates calculated per 100,000 persons. The weight of an age group is the number of persons in it compared to the standard population as per the World Health Organisation (WHO) [24]. The method of age standardization is utilised for an unbiased comparison between populations of different age profiles [7].

Table 2: Top 10 incidence rates of breast cancer globally in the year 2020 (source: WCRF [25]).

Rank	**Country**	**Incidence rate**	
		Number	ASR per 100,000
	World	2,261,419	47.8
1	Belgium	11,734	113.2
2	The Netherlands	15,725	100.9
3	Luxembourg	497	99.8
4	France	58,083	99.1
5	France, New Caledonia	185	99.0
6	Denmark	5,083 98.4	98.0
7	Australia	19,617	96.0
8	New Zealand	3,660	93.0
9	Finland	5,228	92.4
10	US	253,465	90.3

2.2 Breast density

It was noted that women with dense breasts are at a greater risk of developing cancer. The risk can be anywhere from 1.8 to 6.0 times for a woman with highly dense breasts compared to a same-age woman with less breast density [8]. Breast density is determined by the ratio of fibrous and glandular tissues over fatty tissues in the breast. The fibrous and glandular tissue is together called fibroglandular tissue. The fibrous tissue works as connective tissue which keeps the breast tissue intact. The glandular tissue also called lobes is responsible for generating milk, which is then carried to the nipple using ducts. The fatty tissue then fills the space left out by all the other tissues [26]. All these tissues are shown in Figure 3.

According to Breast Imaging Reporting and Data System (BI-RADS), breast compositions are divided into four categories based on the amount of fibroglandular tissue (FGT). These are as follows, **a.** almost entirely fat, **b.** scattered areas of fibroglandular density, **c.** heterogeneously dense and **d.** extreme dense. The categories c and d are also known as dense breasts. The density of the breasts is visually determined by a radiologist. It is not recommended to categorise them into percentages or quartiles. However, it might be possible in the future to quantify the density using breast magnetic resonance imaging (MRI) [2]. The fibroglandular tissue emerges white on a mammogram as it absorbs the ionizing radiation (x-rays) during screening, as can be seen in Figure 3. Similarly, the cancerous cells, also appear white on the mammogram. Hence, fibroglandular tissue can hide a tumour, preventing its early detection. This masking effect was quantitatively measured in a study performed on women with a median age of ≤ 56 years. It was observed that density played a role in masking 50% of the cancers which were later identified within 12 months of the negative screening. Further, the same effect was seen in 26% of all cancers identified after more than 12 months. Hence, at least 50% of the mammograms were unable to identify cancerous cells due to dense breasts [27].

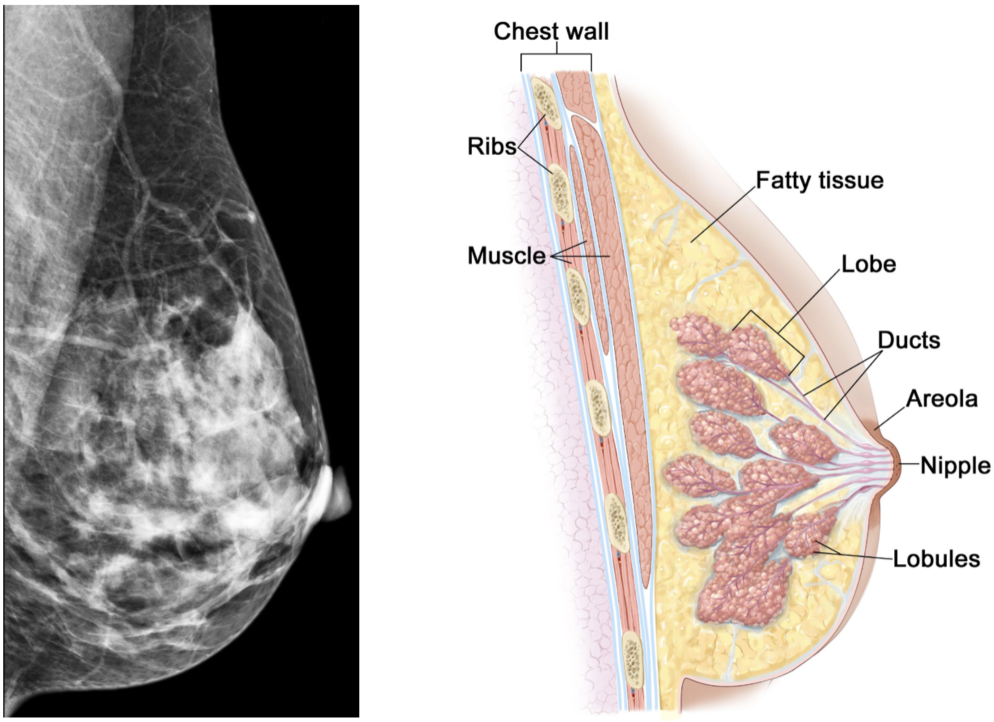


Figure 3: Female mammogram and breast anatomy [left to right]. The white part in the mammogram corresponds to the fibroglandular tissue (lobes, lobules, and ducts). While the grey part represents the fatty tissues (sources: CDC and NCI, USA [6] [26]).

Table 3: Quantitative assessment of mammography to detect tumours in different BI-RADS densities (source: National Center for Biotechnology Information [28]).

Breast density		**Tumours detected**
Category	Meaning	(using mammography)
a	almost entirely fatty	$\approx 100\%$
b	scattered areas of fibroglandular tissue	$\approx 83\%$
c	heterogeneously dense	$\approx 80\%$
d	extremely dense	$\approx 50\%$

Another study was performed to analyse the effectiveness of mammography on breast cancer detection in presence of varying breast density [28]. Table 3 summaries the results of the study. Examining the results it is clear that mammography is most reliable in the case of entirely fatty breasts with 100% cancer detection rate. On the contrary, in the case of extremely dense breasts, mammography only detected 50% of the cancers. Note, it is possible that cancer development occurred during the time between the two screenings. However, the main conclusion remains that it is visually challenging to distinguish between cancerous cells and dense tissue using mammography, as both appear white on a mammogram [28]. In addition to masking cancer, high breast density also increases the risk of developing breast cancer. This risk is approximately 4 to 6 times higher than having almost entirely fatty breasts [27]. In postmenopausal women, breast density was found to be responsible for 26% of all breast cancers [2]. Hence, breast density is considered an important risk factor for breast cancer. It is also an independent risk factor, and if combined with other factors like age and genetics, it could work as a complementary factor. Therefore, the determination of breast density is crucial in providing early detection of cancer and in turn better patient outcomes.

2.3 Breast screening methods

There are several breast screening methods such as ultrasound, mammography, digital breast tomosynthesis (DBT), and magnetic resonance imaging (MRI). The choice of screening method depends on various factors such as cost, availability, and cancer risk. Some of these techniques are discussed further in this sub-section, along with their advantages and disadvantages. Finally, the motivation for using one of them for our study is also described.

1. *Field mammography*

   Mammography is a technique that uses low-dose X-rays to create a black-and-white image of the breast. It is used for various reasons such as screening for breast cancer or determining breast density. The resultant image is called a mammogram or a screening mammogram in asymptomatic women. Screening mammograms often involve two or more X-ray images using different angles from each breast [6]. In case of any symptoms or abnormality in breasts, screening mammograms are used for diagnosing the abnormality, hence referred to as diagnostic mammograms. These often include additional images from different angles, which were not included in the screening mammograms. Consequently, they require higher amounts of radiation. Further, some abnormal-looking areas may be enlarged for better diagnosis. Diagnostic mammograms are also used in cases where the patient has been previously diagnosed with breast cancer [3]. There are various challenges to using mammograms such as false positives, overdiagnosis followed by overtreatment, false negatives, and radiation exposure. Here, a false negative is when an abnormality is initially classified as non-cancerous and later found to be cancer. Further, a false positive denotes a diagnosis that was termed as cancerous but later determined to be not cancer using further screening or biopsy. Overdiagnosis often occurs in cases of ductal carcinoma in situ (DCIS), where suspicious cells start to build up inside the ducts of the breast. It is usually a noninvasive tumour, where these cells would not lead to any symptoms or become lethal [6]. In a 2015 study [29], it was found that mammography has a sensitivity of 86 to 89% in the case of fatty breasts, and 62-68% in the case of extremely dense breasts [2].

2. *Digital mammography*

   Nowadays, digital mammography is more commonly used than field mammography. Similar to field mammography, it generates a breast image using X-rays. However, in this case, the images generated are stored digitally on a computer and not on physical films. The setup is shown in Figure 4. This allows for better management and analysis of these images, as they can be enlarged, edited or annotated. Further, multiple copies can be easily created for this purpose. These can also be very conveniently shared with patients and radiologists [6]. In a

2017 study [30], full-field digital mammography (FFDM) showed a sensitivity of 61% in the biannual screening [2].

3. Digital breast tomosynthesis (DBT)

   It is a kind of digital mammography where X-rays are utilised at various angles to obtain multiple slices of images from the breast. These are then used to construct a 3-dimensional image of the breast using computer software. Hence DBT is also known as 3-dimensional mammography. A computed tomography (CT) scanner also works similarly. Since DBT is often performed simultaneously with 2D digital mammography, the amount of radiation is higher than in standard mammography. However, there are new approaches being researched which allow DBT to be performed without 2D mammography. This has the potential to reduce the radiation dosage [6]. Finally, DBT is widely used for the detection of cancer as it detects more breast cancers. Due to its accuracy, it reduces the need for follow-ups after a screening. Further, it was observed in various studies that it is more effective in the case of dense breasts than other mammography techniques [3]. However, more research is needed to prove its effectiveness in the early detection of cancer and avoidance of false positives, compared to 2D mammography [6].

4. *Breast magnetic resonance imaging (MRI)*

   Another imaging modality used in breast screening is magnetic resonance imaging (MRI). It creates a 3-dimensional digital image of the breast using radio waves and a strong magnetic field. The breast MRI machine is shown in Figure 4. Often times a contrast agent is also used to enhance the quality of the image. It is not recommended as the primary breast screening due to some false negatives and more false positives. This means that a breast MRI can skip some cancers and wrongly flag some masses as cancers. Hence it is not used for women with an average risk of breast cancer development [3]. However, it is used as an additional screening method for patients with high risk such as extremely dense breasts, symptomatic patients or patients with a history of breast cancer [6]. It is also used to monitor breast cancer and determine its exact location, shape and size [3].

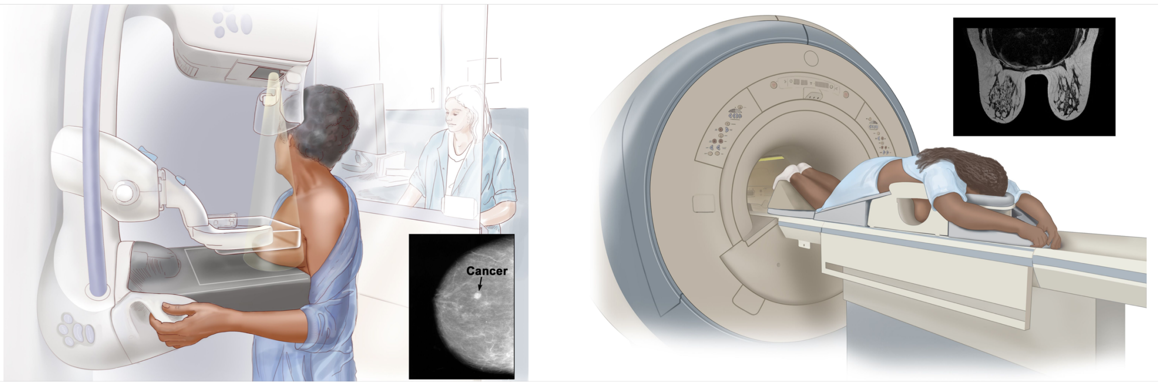


Figure 4: Mammography and breast magnetic resonance imaging [left to right] (source: NCI [6]).

5. *Breast ultrasound*

   It is an imaging technique that uses high-frequency sound waves and their echoes to generate digital images [3]. It is very popular in the medical field as it gives a real-time visualisation, is non-invasive (absence of ionizing radiation like x-rays) and is relatively more economical than other modalities such as MRI [31]. It is not commonly used for screening for breast cancer. However, it can be used in diagnosing some breast abnormalities, for instance, lumps or cysts filled with fluid, which are not distinguishable on a mammogram. Similarly, ultrasound can be very useful in cases of high breast density which also makes mammograms less effective [3]. Ultrasound images also have disadvantages such as low contrast, low signal-to-noise ratio (SNR), and unclear boundaries. Moreover, the presence of speckle noise further deteriorates the quality of these images [32].

A 2005 study compared the performance of three imaging modalities, i.e. mammography, breast ultrasound and breast MRI. They monitored 529 asymptomatic women with an increased risk of $\geq 20\%$ for breast cancer development. This was mostly due to the presence of a breast cancer susceptibility gene (BRCA). The results are summarised in Table 4. It was observed that the least sensitivity was obtained using mammography at only 33%, followed by ultrasound at 40% and their combination at 49%. In contrast, a significantly high sensitivity of 91% was observed using MRI. Further, in higher-risk groups, the sensitivity of MRI was 100%, while for mammography it was 25% which was even less than before. However, the specificities were comparable, with MRI at 97.2% and mammography at 96.8%. Therefore, the following conclusions were drawn for these women with high breast cancer risk due to family history. The use of mammography alone or in combination with breast ultrasound is inadequate for diagnosing breast cancer. In contrast, MRI is much more effective in diagnosing breast cancer at an earlier stage, especially for women with a high risk of breast cancer development [33].

Table 4: Performance of different imaging modalities on breast cancer detection (source: [33]).

Screening method	**Results**		
	Cancers	Sensitivity	Specificity
Mammography	14/43	32.6	96.8
Ultrasound	17/43	39.5	90.5
MRI	39/43	90.7	97.2
Mammography + MRI	40/43	93.0	96.1
Mammography + Ultrasound	21/43	48.8	89.0
High risk groups (21-40%)			
Mammography	5/20	25.0	97.4
Ultrasound	6/20	30.0	91.2
MRI	20/20	100.0	97.7
Mammography + MRI	20/20	100.0	97.0
Mammography and Ultrasound	9/20	45.0	89.9

There has been immense progress in image modalities however, the issue of underdiagnosis is still prevalent. In fact, out of 1000 screenings 8 deaths are prevented, while 11 women die from breast cancer [34]. The underdiagnosis is largely contributed by breast density which in turn affects almost all X-ray-based imaging techniques such as standard mammography, FFDM and DBT [2]. To prevent some underdiagnosis, a breast MRI is often recommended as additional screening, especially in cases of extremely dense breasts [35] [2].

3 Related works

Breast density is commonly classified into four Breast Imaging Reporting and Data System (BI-RADS) categories using the visual assessment by an experienced radiologist. However, since it is a subjective categorization method, it varies between radiologists and sometimes from different readings of the same radiologist [11]. Therefore, a more robust method is required to quantify and further classify breast density. The quantitative methods using image processing techniques give an objective result, which is much more consistent, and reproducible compared to subjective visual assessments by radiologists. In these methods, breast images are analyzed using mathematical algorithms to quantify the amount of fibroglandular tissue present in them. This often involves segmenting the fibroglandular tissue using various techniques such as expectation–maximization (EM) used by [13]. They used various image-processing techniques for segmentation. Initially, they segmented the breast from the body by automatically detecting the boundaries of the body with breast and air, using probabilistic atlas-based segmentation. Subsequently, the segmentation of fibroglandular tissue was performed using the EM algorithm [13].

3.1 Deep-learning based methods

Quantitative methods for breast density estimation have some limitations, such as having a low agreement with other similar methods, and with the visual assessment performed by the radiologist. Note the assessment by radiologists is often used as the ground truth for prediction tasks in radiography. These issues arise because the percentage of dense tissue present in the breast does not necessarily determine the BI-RADS class, as it provides no information about the distribution of the dense tissue in the breast, i.e. the parenchymal pattern, such as scattered or more densely packed [11]. Therefore, just quantifying the amount of dense tissue is not enough, the underlying visual features from the parenchymal patterns must be learned to better classify breast density into the correct BI-RADS category.

Machine learning is a method that tries to learn meaningful features from images to make predictions about them or classify them. Recently, it has been used for an accurate estimation of breast density as it learns the underlying breast features such as overall shape, texture, and parenchymal pattern among others. Some of these characteristics are even unidentifiable by a human visual assessment however, they might be crucial in assessing breast cancer risk. Many researchers have used machine learning for breast density estimation using mammography [16] [17] [18] [11]. Some of the studies have used machine learning on MRI, such as [9], where they used several image segmentation techniques for separating the dense tissue for assessment. First, they segmented lung tissue from the air using fuzzy c-Means (FCM) classification. Second, the chest wall muscle was removed from the images using B-spline curve fitting. Third, the skin was segmented out using dynamic searching. Finally, fibroglandular tissue was segmented from the breast using adaptive fuzzy c-Means (FCM) classification [9].

Recently, more advanced methods such as deep convolutional neural networks (CNN) were introduced to help classify breast density. The ResNet-18 residual network [19], which is widely used for image classification, has been employed for medical image analysis recently [17] [18]. However, most of the previously done studies targeted mammograms instead of MRI images, such as [5], [11], [17], [18] and [16]. MRI is preferred over mammography for breast density estimation as mammograms can have higher variability due to changing body positions, compression levels, and/or X-ray intensities [9]. The studies performed on breast MRI images were very different from our study. Many of them used quantitative methods along with segmentation such as [9] and [13]. Some used regression to estimate breast density [14]. To our best knowledge, there is no recent study that classified breast density using deep learning on breast MRI images without using segmentation or regression. Some of these research papers are summarised in Table 5. The ‘data’ column shows the number of breast scans. Some of these are unilateral, however, for simplicity, this information is skipped.

Table 5: The state-of-the-art research papers for breast density estimation and classification.

Approach	Paper	Method	Data	Results
		Using mammograms		
DL	Lehman2019 [17]	ResNet-18, along with transfer learning	41,479 (train) 8,000 (validate) 8,677 (test)	$\kappa = 0.67$ (initial radiologist), $\kappa = 0.78$ (consensus of radiologists)
DL	Mohamed2018 [18]	AlexNet modified (5 convolutional, 2 fully connected layers), also with transfer learning	22,000 (total) 14,000 (train) 1,850 (test)	$AUC = 0.9857$, $AUC = 0.9882$ (pretrained)
DL*	Saffari2020 [16]	U-net (with skip connections) and pretrained ResNet-101. Cycle GAN (for segmentation)	410 (total) 328 (train) 82 (test)	$Precision = 0.97$, $Recall = 0.97$, $TNR = 0.99$

ML*	Gubern2014 [13]	Atlas-based body-breast segmentation^1, EM-based dense tissue segmentation^2	$(99+1)^1$
			$(26+1)^2$
			(train + test) leave-one-out approach
				DSC$= 0.94 \pm 0.04$,
				Overlap$= 0.96 \pm 0.02$
				(both average)

		Using breast MRI scans		
DL	Velden2020 [14]	3D regression CNN (5 convolutional, 2 fully connected layers)	400 (train) 50 (validate) 165 (test)	$\rho = 0.81$ (spearman’s correlation)
ML*	Nie2008 [9]	Fuzzy c-Means clustering, B-spline curve fitting (segmentation)	600 (total) 11 (train, test)	$\sigma = 3\% - 4\%$ (average between breast volume & percent density)

‘’ denotes the use of segmentation*

Using machine learning or deep learning especially convolutional neural networks (CNN), many researchers have obtained a much higher classification performance than the conventional methods. This has increased the interest in the community to delve deeper into using deep learning in analysing medical images [11]. However, the scarcity of labelled medical data is a great challenge in using deep learning, as it requires a large corpus of data to effectively train a deep learning model. Transfer learning is often used to overcome this issue by using models which are pretrained on some image dataset, such as ImageNet [20]. In conclusion, the search for an optimal deep-learning approach to classify breast density is still ongoing.

4 Methodology

In this section, our proposed methodology is described for breast MRI classification. Further, the characteristics of the data used are discussed, and various experiments and evaluation methods are described. This study aims to classify breast density into four categories and analyze the effects of using different architectures. The pipeline includes four major steps, i.e. preprocessing, model selection, classification and performance evaluation. In preprocessing various techniques were used such as normalisation, resampling, augmentation, cross-validation, and class weighting. In model selection, one of the four models was chosen at a time for training. The models were, 3D ResNet-18 with 3 slices, 3D ResNet-18 with the whole volume, 2D ResNet-18 with ImageNet weights, and 2D ResNet-18 with ImageNet and class weights. Thereafter, a four-class image classification was performed using the selected model. In performance evaluation, the classification performance of each model was compared using ROC (receiver operating characteristic) curves and other classification metrics such as Precision, Recall, F1 score, and Accuracy.

4.1 Dataset

The original dataset consisted of 1627 breast scans from 622 patients, which included both MRI and mammograms. However, the data contained some missing labels. Hence, only 960 MRI scans from 508 patients were taken into consideration. Further, non-fat suppressed T1-weighted MRI sequences were selected from the MRI data. This choice was motivated by the presence of high contrast among fibroglandular and fatty tissues in these images, which is desirable in the case of breast density estimation [15].
The mean age of the patients was 45 years (starting from 19 to 79 years). The distribution of breast density in the dataset is shown in Figure 5. It shows that most of the data were from classes B and C, i.e. scattered fibroglandular density and heterogeneously dense breasts.

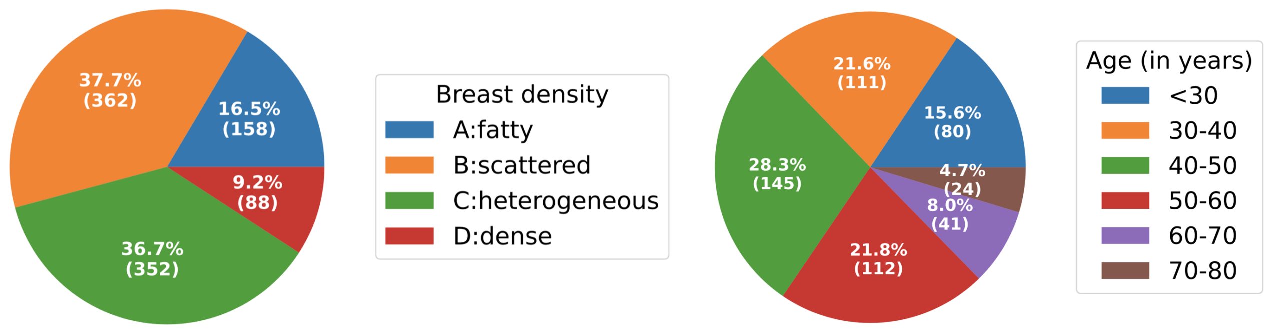


Figure 5: Characteristics of the breast MRI dataset.

4.2 Data preprocessing

Data preprocessing is the cleaning of data by manipulating and/or dropping some data. In most data collection methods, the resultant data is often noisy (i.e. irrelevant, redundant, and/or missing some parts). It is significantly challenging and inefficient to extract knowledge from noisy data, especially in the case of computational biology (i.e. using data analysis to understand biological processes) [36]. Hence, a data preprocessing step is involved to ensure an efficient knowledge discovery from the clean and processed data. Digital image processing is a type of data preprocessing technique which is specifically used for image datasets. It uses a set of computer algorithms to process the images to augment the images to enhance, restore, encode or compress information in them [37]. Any image dataset involves images of varying properties such as pixel intensity values, and image sizes. For a fair analysis of these images, it is often required that some of these properties (such as image size) are standardised across the dataset. In any computer vision application, such as ours, this becomes a crucial step, as the neural networks are designed to take input (in this case images) of a particular size. For our breast MRI dataset, various image processing techniques were used such as conversion from DICOM to NIfTI standard, normalisation, resampling, augmentation, cross-validation, and class weighting. These are explained in more detail further in the section.

4.2.1 DICOM to NIfTI

Two file formats are regularly used for medical images, DICOM and NIfTI. They are short for 'Digital Imaging and Communications in Medicine' [38], and 'Neuroimaging Informatics Technology Initiative' [39]. DICOM is widely used in the medical field due to its ability to store more information and its robustness. However, NIfTI is gaining popularity among data analysts and machine learning applications because it simplifies the handling and visualization of 3D images. DICOM stores each 2D image slice as a separate file, along with metadata such as patient information, acquisition parameters, and image orientation. This can result in a large number of individual files for a single 3D image volume, which can make handling and processing these files more complex. In contrast, NIfTI stores the entire 3D image volume as a single file, along with a smaller set of metadata that includes information such as the image dimensions, voxel size, and data type. This simplifies the handling and processing of the image data, as there are fewer files to manage and the image volume can be easily loaded into memory as a single object. This difference in file format can have important implications for image analysis tasks, such as image processing and classification, which require the processing of large volumes of image data. The more efficient storage and handling of 3D images in NIfTI format can make these tasks faster and easier to perform. Therefore, in our pipeline, the breast MRI scans were converted from DICOM to NIfTI as part of the image preprocessing.

4.2.2 Normalisation

In image processing, pixel normalisation is often done to change the varying range of pixel intensity values to a consistent range (usually 0 to 255). This usually helps the network to learn quickly as the gradients change uniformly. It is done by performing mean subtraction (i.e. subtracting the mean pixel value from all pixel values) to centre the data around the mean pixel value. Further, multiplication by 255 was performed to have a consistent range of (0 to 255) for all the pixel values.

4.2.3 Resampling

In computer vision, resampling or resizing is a crucial step since deep learning models train faster on smaller images. Essentially, the training time of the neural network will increase drastically if it has to learn from four times as many pixels [40]. In addition to that it is also used as a standardisation practice, as neural networks need to be fed same-size inputs, and all images need to be resized to one size before feeding them to a CNN.

- *Downsampling vs upsampling*
  To resize, there are two approaches, downsampling and upsampling. In the downsampling

technique, the resolution of the input image is reduced while trying to preserve most of the information. It results in reduced storage size with the cost of potential deformation of the features, leading to possible degradation in classification accuracy. In upsampling, the reverse is done and try to obtain a higher-resolution output image from the input image by using various methods of interpolation [41]. However, having large images increases both the space and time complexity by occupying more space and increasing the training time of the neural network [42]. As previously seen, there are downsides to both approaches. Hence, choosing a fixed size for images is a tradeoff between computational efficiency and accuracy. Due to restrictions on memory and training time during our experiments, downsampling was performed for resizing our images.

- *Performing downsampling*
  It can be done by cropping or scaling, both are lossy methods. In cropping, the border pixels are removed, thus leading to missing features from the border areas. In scaling, pixels are interpolated to a desired height and width, while choosing to preserve the aspect ratio (width/height). Scaling brings the risk of deforming features or patterns across the image. However, since deformed patterns are still better than missing patterns, scaling is usually preferred over cropping for resizing images down to a desired size [42].

- *Performing downscaling*
  It can be performed using multiple interpolation methods, such as the nearest neighbour, bilinear, and bicubic. Bilinear interpolation is preferred for continuous datasets without distinct boundaries, where the closest points are related. Hence for our dataset, the bilinear interpolation was used since the overall area of white vs grey pixels was needed, and not just their boundaries. To implement this the resize function of the OpenCV library [43] was used. Except for the 3D ResNet-18 model, where 3D scaling was required. Therefore, spline interpolation was chosen for that, as it gives the best trade-off between accuracy and computational cost [44]. To implement this the zoom function of the SciPy ndimage library [45] was used. The final image sizes were as follows, $(512 \times 512 \times 3 \times 1)$ for the 3D ResNet-18 with the three slices; $(128 \times 128 \times 72 \times 1)$ for 3D ResNet-18 and $(128 \times 128 \times 3)$ for 2D ResNet-18 and 2D ResNet-18 with class weights.

4.2.4 Augmentation

Data augmentation is the process of augmenting a dataset with artificial samples created by modifying some of the preexisting data samples. In the scarcity of data, augmentation is often required to enhance the performance of the classifiers. In computer vision, this is often done using image processing techniques such as flipping, rotating, cropping, scaling, adding noise, contrast or brightness, and blurring. Many of these techniques were applied using the imgaug library (version 0.4.0) to augment the dataset [46].

First, the input images were randomly flipped either left or right. Further, randomly cropped was performed from each side by 0 to 0.1 pixels. Half of the input images were randomly selected for blurring using a Gaussian kernel (or a normal function). For every image, the standard deviation (σ) of the kernel was sampled uniformly from the interval [0, 0.1]. More σ means more blurring as it will further spread the distribution, making it less spiky The common values are in the range of 0.0 (no blur) to 3.0 (strong blur). The contrast of the input images was adjusted by scaling each pixel according to the formula: $127 + \alpha \times (v - 127)$ where v is the pixel value and once per image the α is sampled uniformly from the interval [0.90, 1.3]. The multiplier α linearly pronounces (> 1.0), dampens (0.0 to 1.0) or inverts (< 0.0) the difference between each pixel value and the centre value of the data type, e.g. 127 for uint8. A white Gaussian noise was added to the input images, which was sampled once per pixel from a normal distribution denoted by $N(0,s)$, where s is sampled per image and varies between 0 to 0.05×255. Every pixel of the input image was multiplied by a value between 0.8 and 1.2 (i.e. ± 20%) to change the brightness of the image, i.e. to make it brighter or darker.

Affine transformations were applied such as scaling, translation and shear stress. The input images were scaled or zoomed in and out independently per axis by a percentage value of the original size. In our case, this value is sampled uniformly between 80 to 120% per image. Further, the input images are translated to $\pm 20\%$ on the x- and y-axis independently. They are also rotated and shear mapped by $\pm 5°$. The sample images generated after applying the above techniques are shown in Figure 6.

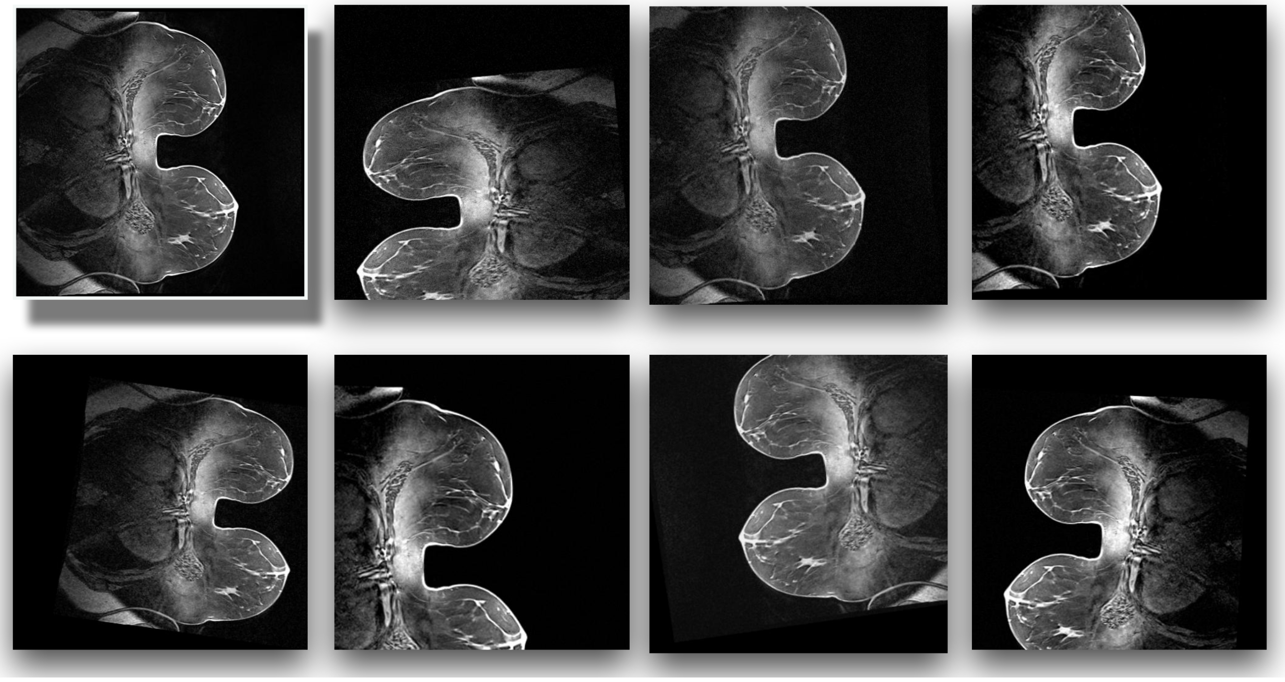


Figure 6: Some sample images were created using different image processing techniques on the original image (top left). The dataset was augmented with a large portion of similar images during training, to solve the issue of data scarcity and overfitting.

4.2.5 Handling class imbalance

Class imbalance is a common problem in machine learning, where the dataset is skewed towards some of the classes. The over-represented classes are called the majority classes; while the under-represented ones are known as the minority classes. As a result of this imbalance in the dataset, the classifier becomes biased towards the majority classes and performs poorly on the minority classes [47]

There are various ways to address this issue, such as undersampling the majority classes, oversampling the minority classes, an ensemble of both, or class weighting. Since our dataset is already small, undersampling is not ideal as that would further decrease the amount of training set. Oversampling however can be favourable for us, as it involves creating more samples of the minority classes. There are various ways to perform oversampling, the two most common ones are Random Oversampling, and Synthetic Minority Oversampling Technique (SMOTE) [48]. Introducing class weights or balanced weights is another useful technique to handle class imbalance. Although, it does not involve re-sampling, instead works with the prioritization of the existing samples. In other words, it gives higher weights or priority to the under-represented classes during the training phase. Both oversampling and class weights are favourable techniques for our limited dataset. Further in this section, more details are given about these methods and motivate our choice of method.

- *Random oversampling*
  Random Oversampling is among the earliest techniques introduced to handle class imbalance issues. It became popular because of its robustness. It works by augmenting the dataset with multiple copies of randomly selected samples of the minority classes. It can be performed multiple times to create a roughly balanced dataset. However, creating multiple copies of some of the minority class samples can cause the classifier to memorize those samples, instead of learning the underlying patterns. This could lead to overfitting and poor performance of the model in practice [49].

- *SMOTE*
  Synthetic Minority Oversampling Technique was introduced to overcome the challenge of overfitting in the random oversampling technique. It creates synthetic minority class samples using a subset of the existing samples chosen randomly with replacement. However, unlike random oversampling, it broadens the decision region by operating in the 'feature space' rather than 'data space' [50]. More specifically, new synthetic examples are only added along the line segments joining some or all of the k nearest neighbours of the minority class starting from a given sample of the minority class [51]. There are some limitations to SMOTE, some of which are discussed here. Firstly, as with the use of any oversampling method, the classifier overestimates the probability of a sample belonging to the minority class, which may lead to a poorly evaluated model. Secondly, SMOTE may produce noise while generating synthetic samples, as it only considers the minority-class neighbours and disregards all the other classes in the neighbourhood of a given minority-class data point. Thirdly, it can only generate new samples inside the range of the existing minority samples. Therefore, in the case of a sparse minority class dataset, it might not be effective in generating enough synthetic samples to create a balanced dataset. Additionally, SMOTE is not very effective on higher dimensional data. As with increasing dimensions, it becomes more and more challenging to find out the nearest neighbours of a sample. Hence, it could lead to inaccurate synthetic samples. Lastly, SMOTE may not be suitable for an imbalanced dataset where the minority class is located in dense regions of the feature space. This is because the synthetic samples generated by SMOTE may not be representative of the true underlying distribution of the minority class, leading to poor classification performance. Therefore, in such cases more advanced techniques, for instance, cluster-based oversampling or adaptive synthetic sampling are preferred over SMOTE.

- *Class weights*
  Class weighting is a technique used to assign higher weights to the minority class and lower weights to the majority class. This gives the model a better understanding of the importance of each class and helps it to make better predictions for the minority class. Class weighting adjusts the cost function of the model so that misclassifying an observation from the minority class is more heavily penalized than misclassifying an observation from the majority class. This approach can help to improve the accuracy of the model by rebalancing the class distribution.

4.3 Image classification

The process of automatic image classification is done in a supervised learning approach, where the images and their true class labels are given to a machine learning or deep learning model. The output of the model is the predicted class labels. These image classifier models extract meaningful information from the images such as pixel intensity values concerning their local area, and edges. The mathematics behind this decision-making process of a machine learning model is explained further in this section. Various deep architectures have been derived from traditional feed-forward artificial neural networks (ANN). It consists of a cascade of trainable multi-stage layers inspired by the organisation of the animal visual cortex [52]. There are sets of arrays called feature maps as the input and output of each layer. Each feature map in a specific layer represents particular features extracted at the locations of the associated input.

4.3.1 Neural networks

An artificial neural network (ANN) as the name suggests is a network of artificial neurons. Here, neurons are the computing units, which are connected to each other in multiple layers. These neurons take input and give output using the incoming and outgoing connections. These connections have a scalar weight associated with them [53]. Using the input, the neurons execute some calculations with the help of an activation function and give the output, which can be used by neurons of the next layer. The layers are categorised into three types, these are, an input layer which receives the input from the user, one or more hidden layers, which perform the internal calculations and finally an output layer, which gives the output of those computations. Further, the computations or learning can be improved by altering the weights (or parameters) of the connections. The artificial neural network can be represented as a function $y = f(x \times q)$ which maps the input values x to the output values y using parameters q. The aim is to approximate $y = f(x)$, and parameters are selected by minimising the error associated with the approximation. An error function is used for computing these errors with the help of derivatives of the function with respect to the parameters.

4.3.2 Gradient descent

A gradient is a vector containing all partial derivatives of a function g(x,y). In other words, it is a vector field describing the slope of the tangent pointing in the direction of the largest rate of increase of the value of the function [54]. Gradient descent is an optimization algorithm which is used in machine learning and deep learning. It is used to minimize the cost function of a model by adjusting the parameters iteratively in the direction of the steepest descent in each sweep [55]. It works by computing the gradient or change of the cost function with respect to all the parameters and then updating each parameter proportional to its gradient. This change in the value of the parameter happens while back-propagating the information in an ANN. The gradient value calculated from the backpropagation is used to update the weights as shown in Equation 1. Here, ∇F represents the gradient, $\gamma \in [0,1]$ is the learning rate of the network, $\mathbf{a}$ is the parameter of the function. The parameters are updated until the algorithm converges to a local minimum.

$$\mathbf{a}_{n+1} = \mathbf{a}_n - \gamma \nabla F(\mathbf{a}_n) \tag{1}$$

The amount of shift of the parameters in the direction of the negative gradient is determined by the learning rate γ. It is an important hyperparameter as it determines how fast or slow the network will learn. A high learning rate requires a lesser number of training iterations.

4.3.3 Stochastic gradient descent

Stochastic gradient descent (SGD) is a kind of gradient descent in which a subset of training examples are randomly selected to calculate the gradient at each iteration. In other words, it does not compute the gradient for each training sample separately and instead uses a mini-batch of samples. This makes it computationally more efficient than standard gradient descent especially in the case of large datasets. However, there are some drawbacks to using SGD such as noise and less accuracy than standard gradient descent.

4.3.4 Loss functions

A loss function maps the values of some variables associated with an event, to a real number which represents the cost in relation to the event. Consequently, it is also known as an error function or cost function [56]. For classification tasks, the majority of researchers apply the logarithmic (or cross-entropy) loss to the output of the network which in turn uses softmax activation in most classification cases. The formula for the logarithmic cross-entropy loss is shown in Equation 2. Note, here y denotes the true one-hot encoded label, **o** is the output of the last layer of the network, $.^{j}$ denotes the j^{th} dimension of the given vector, and $\sigma(.)$ denotes the estimated probability [57].

$$Cross-entropy\ loss = -\sum_{j} y^{(j)} log\ \sigma(\mathbf{o})^{(j)} \tag{2}$$

The one-hot encoded labels are used in the case of categorical cross-entropy. Sparse categorical cross-entropy is used in the case of two or more label classes where the labels are in the form of integers instead of one-hot encoded.

4.3.5 Activation functions

Activation functions are mathematical functions that are used in ANN to introduce non-linearity into the output of a neuron. They are applied to the weighted sum of the inputs to a neuron and produce an output that is passed on to the next layer of the network. Activation functions are an essential component of neural networks because they allow the network to learn complex non-linear relationships between the input and output of a network.

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{3}$$

A softmax activation function is often used when a normalized probability distribution is required as the output. In other words, the output of the softmax function is in the interval (0,1) and sums up to 1. Before using softmax, it is possible that some of the vector components were greater than one or negative. The equation is shown in 3, where, $\vec{z}$ is input to the softmax function in the form of a vector (z_0, ... z_K), with z_i denoting the elements of the vector, each taking a real value. Further, e^{z_j} denotes a standard exponential function, which is applied to every element of the input vector z_j. K denotes the total number of classes or labels. The lower part of the fraction is the normalization term, which makes sure that the output of the softmax sums up to 1 and is in the range (0, 1).

4.3.6 Convolutional neural networks

In a convolutional neural network (CNN), kernel convolution operations replace the general matrix multiplications that are seen in feed-forward artificial neural networks [58]. This is because CNN is designed to work with input data that has a grid-like structure, such as images or audio signals. The kernel convolution operation involves sliding a small matrix (the kernel) over the input data and computing the dot product between the kernel and the input data at each position. This produces a feature map that captures local patterns in the input data. By stacking multiple convolutional layers, CNN can learn increasingly complex features from the input data.

In a CNN, a convolution operation is used to detect the patterns in the input images. These patterns are crucial to predict the correct label of images. Additionally, it is also commonly used in image processing. A convolution is an element-wise multiplication using a sliding kernel on the input image $f(x,y)$, that gives a feature map $g(x,y)$ as the output [59]. The 2-dimensional and 3-dimensional convolution kernel operations are shown in Equation 4 and 5 respectively.

$$g(i,j) = \omega * f(x,y) = \sum_{dx=-a}^{a} \sum_{dy=-b}^{b} \omega(dx,dy) f(x-dx, y-dy) \tag{4}$$

$$g(i,j,k) = \omega * f(x,y,z) = \sum_{dx=-a}^{a} \sum_{dy=-b}^{b} \sum_{dz=-c}^{c} \omega(dx,dy,dz) f(x-dx, y-dy, z-dz) \tag{5}$$

Note in these equations, $g(i,j)$ is the output feature map, $f(x,y)$ is the given image with x and y as its indices, ω is the kernel matrix for convolution with dx, dy as its indices, and $*$ is the convolution operation. An example of the 2-dimensional convolution operation is shown in Equation 7 which uses the matrices shown in Equation 6. Further, an example of the 3-dimensional convolution operation is shown in Figure 7.

$$f(x,y) = \begin{bmatrix} 3 & 3 & 2 & 1 & 0 \\ 0 & 0 & 1 & 3 & 1 \\ 3 & 1 & 2 & 2 & 3 \\ 2 & 0 & 0 & 2 & 2 \\ 2 & 0 & 0 & 0 & 1 \end{bmatrix}, \omega = \begin{bmatrix} 0 & 1 & 2 \\ 2 & 2 & 0 \\ 0 & 1 & 2 \end{bmatrix} \tag{6}$$

$$g(i,j) = \begin{bmatrix} \mathbf{3} & \mathbf{3} & \mathbf{2} & 1 & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & 3 & 1 \\ \mathbf{3} & \mathbf{1} & \mathbf{2} & 2 & 3 \\ 2 & 0 & 0 & 2 & 2 \\ 2 & 0 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} 0 & 1 & 2 \\ 2 & 2 & 0 \\ 0 & 1 & 2 \end{bmatrix} = \begin{bmatrix} \mathbf{12.0} & 12.0 & 17.0 \\ 10.0 & 17.0 & 19.0 \\ 9.0 & 6.0 & 14.0 \end{bmatrix} \tag{7}$$

Concerning deep learning, a convolution kernel helps the network to learn low and high-level features as the information propagates in the network from input to output.

4.3.7 Residual neural network (ResNet-18)

A residual network (ResNet) is a kind of convolutional neural network which was first introduced by Kaiming He et al. in 2015 [60]. The residual neural network has shown state-of-the-art performance on several image classification tasks. Consequently, it is being used in various applications concerning images such as object detection and semantic segmentation. Recently, many researchers are using the residual network in medical imaging for the analysis of medical images such as automatic diagnostics [61]. The residual neural network is considered to be a directed acyclic graph (DAG) network because of its complex layered architecture, in which the layers give and receive input from

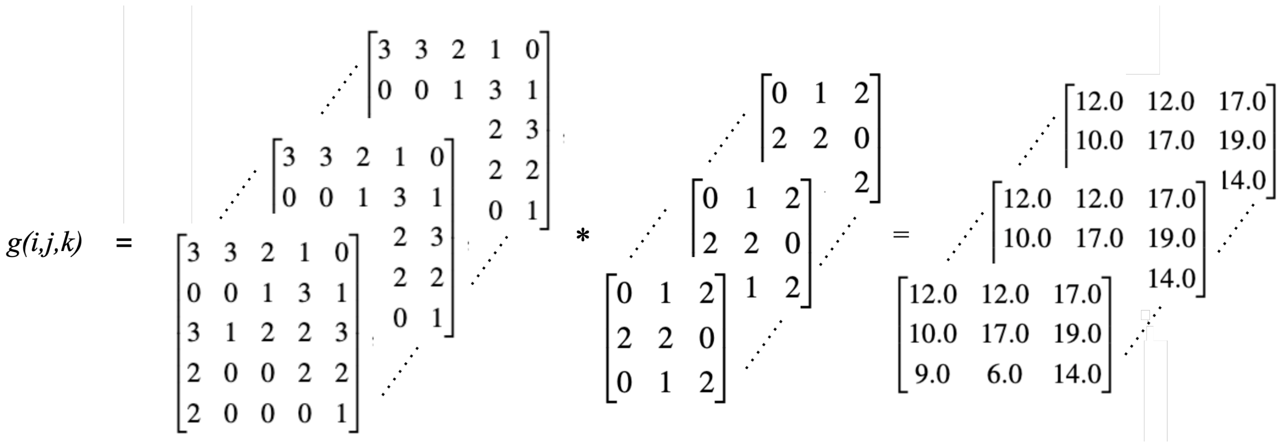


Figure 7: Three-dimensional convolution operation using 3D kernel on a 3D input image to preserve the information stored in the third dimension, resulting in a 3D feature map as the output.

multiple layers. It has many variants such as ResNet16, ResNet18, ResNet34, ResNet50, ResNet101, ResNet110, ResNet152, ResNet164, and ResNet1202 depending on the number of deep layers. For instance, the ResNet18 has 72 total layers along with 18 fully connected or deep layers. The main idea behind a ResNet is the use of shortcut connections with identity mappings to help resolve the issue of vanishing gradients arising during the training of deep networks through backpropagation. The shortcut connections allow the reuse of activations from previous layers which can help prevent the issue of vanishing gradient. Further, it results in efficient and faster training by allowing them to skip over layers. Residual blocks were introduced as part of the ResNet architecture. These are the skip-connection blocks which allow an easier learning of residual functions in reference to the inputs of the layers, instead of much more complex learning of unreferenced functions.

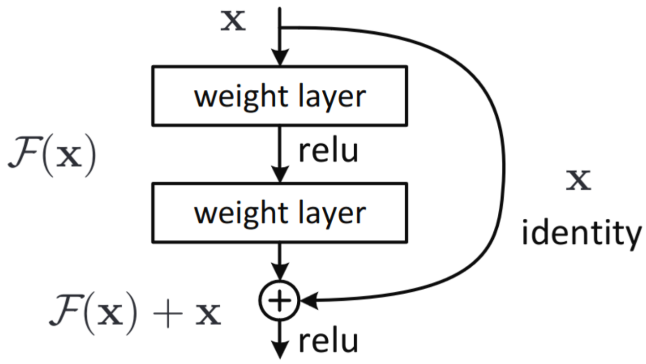

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 8: Residual block in a residual network (ResNet) (source: [60])

The residual block can be seen in Figure 8, where $F(x)$ is the residual function. It helps overcome the issues of vanishing gradients using skip connections and identity mappings. The identity mapping $F(x) = x$, basically gives the same output as the input. Further, using this mapping the output or residuals can be made zero i.e. $F(x) = 0$. This is much more difficult to achieve otherwise using non-linear layers of the network, i.e. $F(x) = x$. Hence, identity mapping can optimally compress the network. Further, in ResNet, the compression of the connections is followed by the expansion of the layers so that the residual part of the network could also train and explore more feature space.

4.3.8 Cross-validation

In machine learning, cross-validation is a technique used to validate the performance of a predictive model on unseen, real-world data. A commonly used cross-validation technique is K-fold cross-validation. It involves dividing the data into k subsets or folds of equal size. One of the folds is held out for validation. The model is trained on k-1 of the other folds (training set) and evaluated on the remaining fold (validation or test set). The process is repeated k times, with each fold being used as the validation set once. An evaluation score is obtained from each fold, their average gives the overall score for the model. The overall score is significantly more reliable than a test score obtained from a single train-test split [62].

- *Overfitting*
  When the size of the training set is small and/or the number of parameters in the model is very large, there is a greater risk of overfitting. Meaning, the model will be overly complex and will fit the noise in the training data rather than the underlying features. These conditions are true for our case, our dataset is fairly small with only 960 MRI scans, and the model parameters are huge, more than 11 million per model. Using cross-validation in this case, a better estimate of the true performance of our model can be obtained, by repeating the training and evaluation process multiple times on different subsets of the data. It also helps us identify overfitting, for instance, if the model performs consistently well across all folds, it suggests that the model is likely to generalize well to new data. However, if the performance varies widely across different folds, that means that the model may be overfitting the training data and might need further regularization to improve its performance.

- *Stratified K-fold Cross-validation*
  A type of cross-validation that creates k-folds while maintaining similar class distribution as the original dataset is called stratified cross-validation. It helps improve and stabilise the performance of the model across folds by preventing the overfitting or underfitting of the minority class. In contrast, if a fold had a significantly different class distribution than the others, it could lead to bias in the performance evaluation. Consequently, stratified cross-validation is a significantly useful resampling and validation technique while working with an imbalanced dataset. This applies to our dataset with two of the classes representing 74% of the data, and the remaining two representing 26% only. Therefore, the stratified k-fold cross-validation technique was chosen for our experiments to preserve the class distribution while training on multiple folds.

5 Experiment setup

This section describes the various setups performed before training the CNN models. First, the deep learning pipeline is discussed along with the architecture diagram of ResNet-18. Then, the technologies and resources used in the study are described. Afterwards, the experiment configurations are discussed which include acquisition parameters for MRI, and hyperparameters for CNN training. Finally, the performance measures are described along with the equations and motivation for each.

5.1 Deep learning pipeline

The entire pipeline of the classification system is shown in Figure 9. It has various steps, divided into two sections, preprocessing along with ImageNet and class weights, and classification. Firstly, the input size is selected as either 3 slices or 72 slices. Note, the latter is also indicated as the entire volume of the MRI image throughout our study. Also note, all these MRI images were already converted from DICOM to NIfTI format before being used as the input. Secondly, various image preprocessing steps were performed on the input images. These include normalisation, resampling (specifically, downscaling), augmentation, and stratified k-fold cross-validation. Augmentation was done using various other image processing techniques, which are shown in Figure 10. These are explained later in the section. Thirdly, transfer learning was applied using the ImageNet weights. Further, class weights were applied to handle the class imbalance problem. Fourthly, the processed images were sent as input to different classifiers such as 2D, 2Dw, 3D3 and 3D CNNs. These are short for 2-dimensional, 2-dimensional class-weighted, 3-dimensional with 3 slices, and 3-dimensional with all (i.e. 72) slices. All of them are based on the architecture of ResNet-18 as shown in Figure 11. Finally, the output of these CNN models was obtained as four predicted probabilities, one for each class being the true class label. The highest probability was chosen as the class label for the input MRI scan.

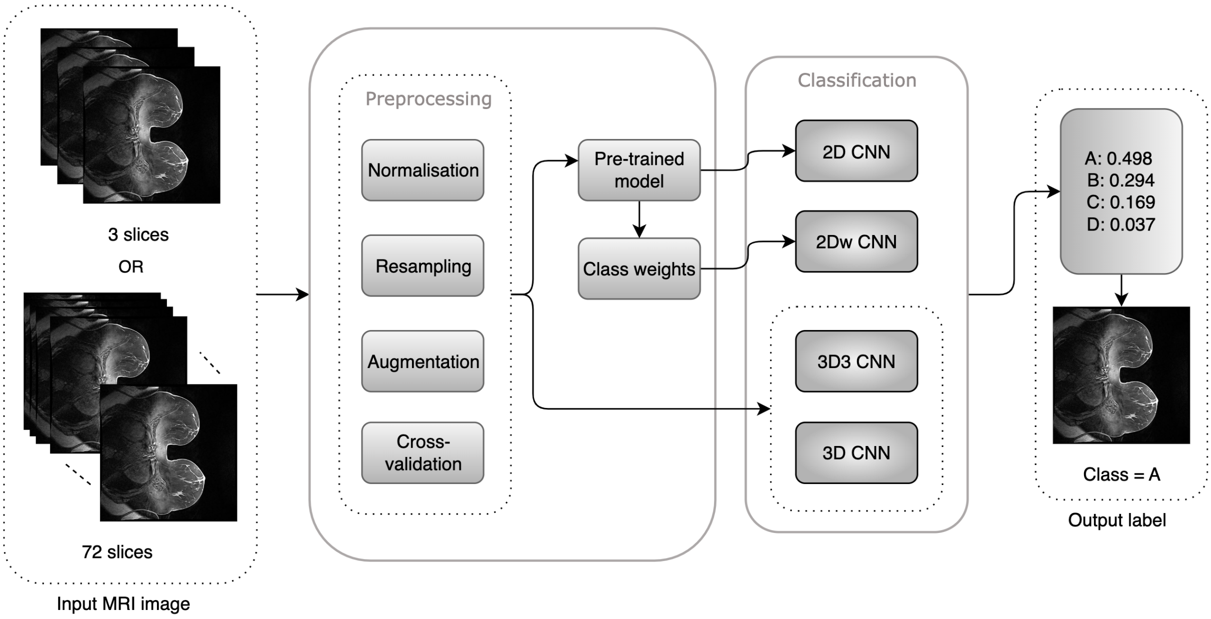


Figure 9: The entire pipeline of our breast-density classification system. Here, A, B, C, and D are the BI-RADS categories for breast density. 2D, 2Dw, 3D3 and 3D CNNs are the adapted CNN models representing 2-dimensional, 2-dimensional class-weighted, 3-dimensional with 3 slices, and 3-dimensional with all slices. All these are based on ResNet-18 architecture.

As discussed previously our classification system included various image processing techniques to augment the dataset as shown in Figure 10. These include flipping, cropping, Gaussian blurring, adding linear contrast, modifying the brightness, adding white Gaussian noise, translating, rotating, and shear mapping the image. The imgaug library (version 0.4.0) was used to apply these techniques for augmentation [46]. Firstly, a horizontal flip was applied to 50% of the input images selected at random. This flipped the images which are stored as a float32 array, horizontally from left to right. Secondly, random cropping was performed on the input images. Each side of the image was cropped individually using a random fraction sampled uniformly from the interval $[0, 0.1]$. The range represents the percentage of height and width of the image. In this case, it indicates a crop of 0 to 10% of the original height and width. Thirdly, half of the input images were randomly blurred using a Gaussian kernel. For every image, the standard deviation (σ) of the kernel was sampled uniformly from the interval $[0, 0.1]$. More σ means more blurring as it further spreads the Gaussian (or normal) distribution. Commonly the value is in the range of $[0.0, 3.0]$, where 3.0 denotes a strong blur. Moreover, the contrast of the input images was adjusted by scaling each pixel according to the formula: $127 + \alpha \times (v - 127)$ where v is the pixel value and α is sampled uniformly from the interval $[0.90, 1.3]$ for each image. The value of α can have a wide effect on the images. It changes the difference between the pixel values and the centre value of the data type of the array in which the images are stored. The centre value in our case is 0 as our images are stored in arrays of float32 data type. If $\alpha > 1.0$, the difference is linearly pronounced, if $\alpha = [0.0, 1.0]$ it is dampened, and if $\alpha < 0.0$ it is inverted.

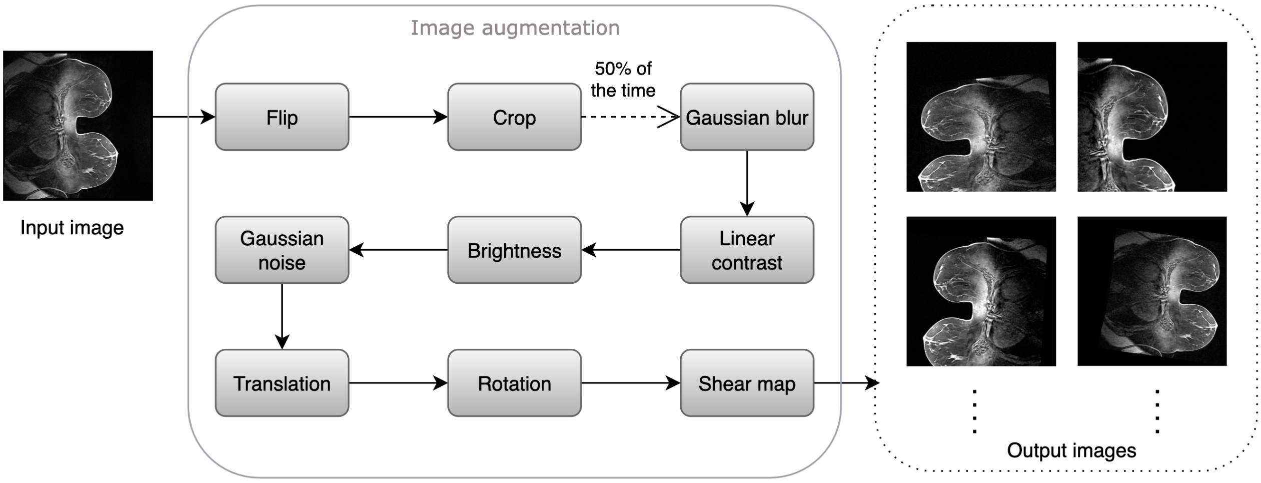


Figure 10: The flowchart showing the augmentation process. It involves various image processing techniques such as flipping the image left to right, cropping it from the sides, blurring it with a Gaussian kernel, applying linear contrast, changing the brightness, adding white Gaussian noise, translating, rotating, and shear mapping the input image.

Additionally, the brightness of the input image was also either increased or decreased. It was done by multiplying every pixel of the image with a value between 0.8 and 1.2 (i.e. ± 20%). Further, a white Gaussian noise was added to the input images, which was sampled once per pixel from a normal distribution $N(0, s)$, where s is sampled per image and varies between 0 to (0.05×255). Finally, different affine transformations were applied such as scaling, translation and shear stress. The input images were scaled in and out independently per axis. The amount of scaling was determined by a percentage value of the original size. In our case, this value was sampled uniformly between 80 to

120% per image. Further, the input images were translated to ±20% on the x and y axes independently. The images were then rotated and shear mapped by ±5°. The sample images generated after applying the above techniques are shown in Figure 6.
Our models were built on the residual network with 18 deep layers (ResNet-18) adapted from [19]. The architecture of the ResNet-18 is shown in Figure 11 where the layers are connected by solid or dotted lines. The former is used when the input and output dimensions are the same. The latter is used when there is an increase in dimensions. Here, identity mapping is still performed however using zeros padding due to increased dimensions along with a stride of 2. In our implementation, some modifications were done to this architecture such as the input layer was slightly modified for different CNN models as shown in Table 6. Further, two additional output layers were introduced. These were a global-average-pooling layer, which was different for 2D and 3D models, and a dense layer. The output of the CNN was four predicted probabilities for four BI-RADS categories. The highest probability gave the predicted class label for the input MRI image. The functioning of the residual block in the ResNet architecture is explained in Section 4.3.7.

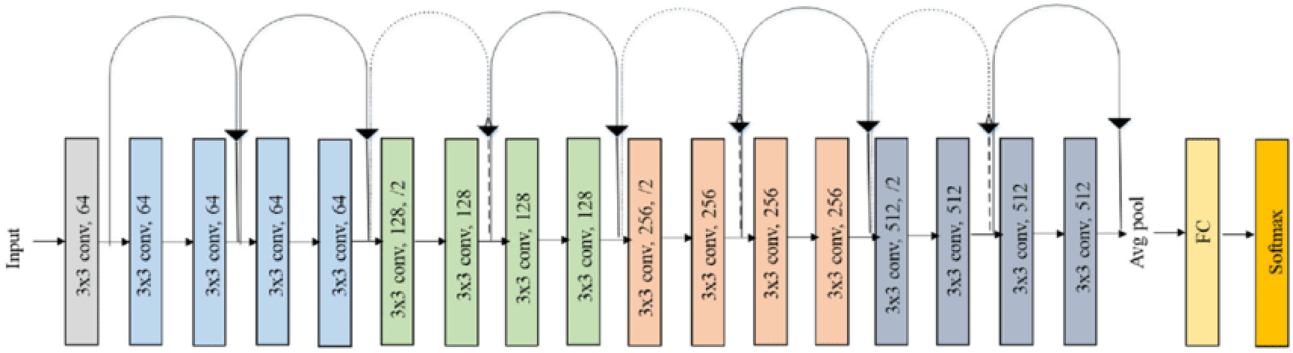


Figure 11: The architecture of ResNet-18, where FC is the fully connected dense layer, Avg pool is the average pooling layer, conv denotes the convolution operation in the corresponding convolution layer and softmax is the activation function. The eighteen layers include 17 convolutional layers with 3x3 filters, the fully-connected (FC) layer and a softmax layer for classification. A stride of 2 is used for downsampling in some convolutional layers. The curved arrows denote the shortcut connections and subsequent residual blocks (source: [63]).

Table 6: Summary of input sizes and the number of parameters for different CNN models.

Model	**Input size**	**Number of parameters**		
		Total	Trainable	Non-trainable
2D CNN	$[128 \times 128 \times 3]$	11,189,025	11,181,083	7,942
2D class weighted	$[128 \times 128 \times 3]$	11,189,025	11,181,083	7,942
3D with three slices	$[512 \times 512 \times 3 \times 1]$	33,172,451	33,164,513	7,938
3D with whole volume	$[128 \times 128 \times 72 \times 1]$	33,172,451	33,164,513	7,938

5.2 Technologies and resources

The magnetic resonance imaging (MRI) examinations were performed on two scanners. These were the 1.5-T (MAGNETOM Skyra, Siemens Healthcare) scanner and the 3.0-T (MAGNETOM Avanto fit, Siemens Healthcare) scanner. A labelled MRI dataset containing 1627 scans from 622 patients is obtained from a radiologist at UMCG. All the experiments were performed on a 64-bit Windows PC provided by the University Medical Center Groningen (UMCG) for the project. Further, Nvidia Quadro P6000 GPU was employed for the training of the CNN models. Python3 [64] was used as the programming language for the project as it supports multiple libraries and frameworks for artificial intelligence such as Scikit-learn, Pandas, Keras, and TensorFlow. Keras library (version 2.10.0) [65] was used as the interface for developing the convolutional neural networks. Further, the TensorFlow framework (version 2.10.0) [66] was used as the backend for Keras. The architecture for the 3-dimensional ResNet-18 models was a modified variant of the 3-dimensional classification models available on the $ZFTurbo/classification\_models\_3D$ GitHub repository [67].

5.3 Experiment configurations

This section describes the different configurations used in the experiment. Starting with the acquisition parameters for the MRI images, followed by the hyperparameters used during model training, and finally, the performance measures to compare models among each other.

5.3.1 MRI acquisition parameters

In the collection of MRI images, two scanners were used, namely 1.5-T and 3.0-T. A total of 423 and 537 MRI scans were obtained from both scanners respectively. Table 7 shows the parameters used while generating MRI scans. Here, TR denotes repetition time, TE denotes echo time, Flip is the flip angle, FOV is the field of view, and Resolution is the in-plane resolution. Both 1.5-T and 3.0-T scanners have the same flip angle and field of view. However, 1.5-T has higher repetition and echo time. On the other hand, 3.0T has a greater number of matrices and in-plane resolution.

Table 7: Acquisition parameters for the breast MRI

Scanner	**TR**	**TE**	**Flip**	**Matrices**	**FOV**	**Resolution**	**Scans**
	(ms)	*(ms)*	*(°)*	*(N × N)*	*(mm)*	*(mm)*	*(N)*
1.5-T	5.27	2.39	10	384 × 338	350 × 350	0.46 × 0.46	423
3.0-T	4.50	1.63	10	416 × 370	350 × 350	0.84 × 0.84	537

5.3.2 Hyperparameters

Several hyper-parameters were studied and selected empirically during initial phase of training. Batch sizes of 8 were chosen after experimenting with 2, 4, and 16 to limit the computational power. The models were trained for 50 epochs. Here, the epoch denotes the number of iterations in which each data sample was used to train the model parameters. Several optimisers were tested; however, stochastic gradient descent with momentum gave the lowest validation loss. The learning rate was set to 0.05 and was selected empirically among the values 0.001 and 0.01. Sparse Categorical Cross entropy was used as the loss function for the breast density prediction as it was a classification task. Sparse Categorical Accuracy was used as the metric. All experiments shared the same training hyperparameter configurations.

5.4 Performance measures

The performance of the models was measured using several evaluation metrics. First, stratified 5-fold cross-validation was applied, where one fold was kept as the held-out validation set, while the model trained on the remaining folds. The trained model was then used to predict the outcomes of the validation set. The validation results were evaluated using multiple classification metrics such as precision, recall (sensitivity), specificity, accuracy, and F1 score for a general measure of performance. Their formulae are described in the equations 8, 9, 10, 11, and 12. Note, here TP, TN, FP, and FN are short for true positives, true negatives, false positives, and false negatives. While PPV and TNR denote positive predictive value and true negative rate. In binary classification, the two classes are called the positive and negative classes. Further in the medical domain, the positive class is often the one indicating a disease or lesion, and the negative class is the absence of the disease. In this study, each BI-RADS category was chosen as the positive class at a time, while the remaining were considered the negative class.

$$Precision(PPV) = \frac{TP}{TP+FP} \tag{8}$$

$$Sensitivity(Recall) = \frac{TP}{TP+FN} \tag{9}$$

$$Specificity(TNR) = \frac{TN}{TN+FN} \tag{10}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{11}$$

$$F1score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{12}$$

Additionally, a confusion matrix or error matrix was used to visually describe the classification performance of all models on different classes. It is commonly used in practice to see how much the classifier is confusing two or more classes. As it has a table-like layout with a horizontal and vertical axis denoting actual or predicted classes or vice versa in some literature [68]. For our case of multiclass classification, the standard confusion matrix needed to be modified to include four classes instead of two. It was a straightforward modification by adding more rows and columns. Further, by considering one class as positive and the remaining as part of the negative class. Furthermore, a receiver operating characteristic (ROC) curve was used for the graphical visualisation of the performance. It plots two metrics, sensitivity and (1-specificity), in other words, true positive rate versus false positive rate. However, since this was a four-class classification task, a normal ROC curve needed to be modified to include more than two classes. Hence, a one-vs-rest multiclass ROC was used, in which each class is taken as the positive class and the others remaining are considered as part of the negative class. Moreover, the area under the curve (AUC) values were analysed for comparison of the performance of each model on each class.

6 Results

This section summarizes the results of the experiments performed during the study. It is divided into two sections, performance of models and cross-validation results. In the former, the overall performance of the CNN models is described using various metrics such as precision, recall, f1 score, and accuracy. For visualisation and further analysis, receiver operating characteristic (ROC) curves are used and their area under the curve (AUC) values are described. Additionally, confusion matrices are drawn for further visualisation of the multi-class classification results of different models. In the second section, the individual results are shown from the 5 folds in the stratified cross-validation method. One model is selected to depict the receiver operating characteristic (ROC) curves for all of its 5 folds. Further, the confusion matrices for the model are shown for each fold. For all these statistical analyses of the results, the scikit-learn library (version 1.0.2) was used [69]. As it is an open-source and efficient method for predictive data analysis, which is built on top of NumPy, SciPy, and Matplotlib libraries.

6.1 Performance of models

The metrics used for the analysis of classification performance were precision, recall, f1 score, and accuracy. The formulae for the same are described in the experiment setup section 5.4. The values of these metrics for different models are summarised in Table 8. Note, these have been rounded to two decimal places for quick interpretation of results. Since it is a multi-class classification task, simple binary metrics had to be modified. For that, the definition of positive and negative classes was modified. one class for a model was taken at a time and considered to be a positive class and the remaining three classes as the negative class. Then, the values of precision, recall, f1 score and accuracy were computed as shown in the equations 8, 9, 10, 11, 12 in section 5.4. This gave four values, one for each density class *(a,b,c,d)* of a model. The overall values were calculated by taking the average of the individual values for each class. Table 8 shows the overall values of precision, recall, f1 score and accuracy for a model.

Table 8: Mean values of various metrics for multi-class classification.

Model	**Precision**	**Recall**	**F1 score**	**Accuracy**
2D ResNet-18	0.59	0.55	0.56	0.78
2D ResNet-18 with class weights	0.65	0.59	0.61	0.83
3D ResNet-18 with three slices	0.56	0.53	0.54	0.75
3D ResNet-18	0.51	0.48	0.48	0.70

In addition to the quantitative analysis, a graphical visualisation of the performance was also included. For this purpose, receiver operating characteristic curves were chosen as they are widely used for performance analysis in the medical domain [70]. The main reason for that is them being uninfluenced by the changes in the decision criteria or threshold. Hence, the area under the curve captures the true capacity of a classifier to discriminate between two classes.

Since ROC curves are designed for binary classification tasks they needed to be modified for multi-class classification. For that, One-vs-Rest multiclass ROC curves are generated by considering one class at a time. Hence, the considered class would act as the positive and all the others as the negative class. This would generate multiple ROC curves for each class, as shown in Figure 12. Note that these graphs contain multiple lines, that is because the ROC curves of different models were combined for a particular class in one graph. The motivation was a clear comparison of different models.

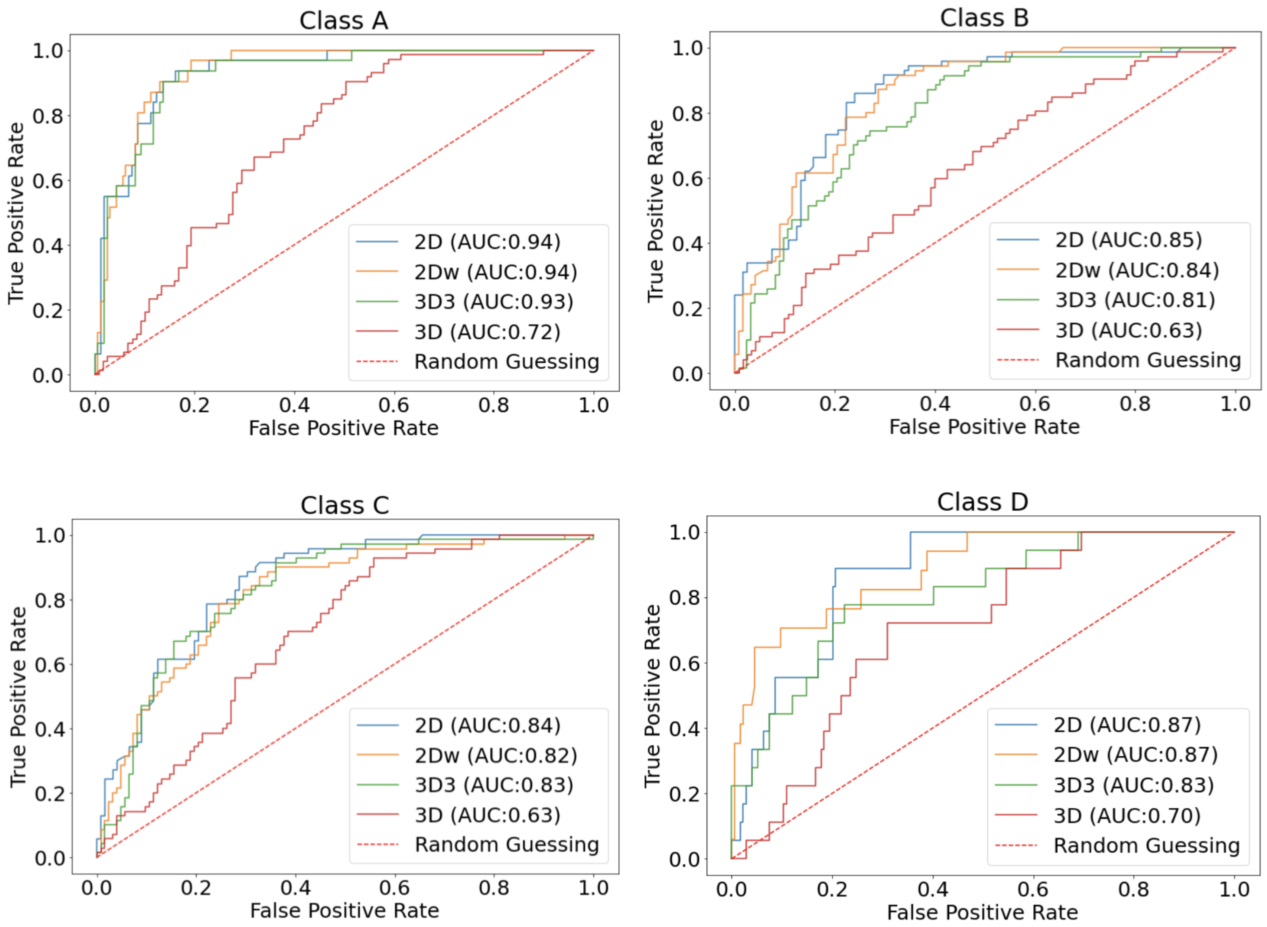


Figure 12: Receiver operating characteristic curves for the multiclass classification on the validation set. The area under the curve for 2D CNN was 0.94,0.85, 0.84 and 0.87 for A, B, C, and D classes respectively. Similar values were depicted for 2D CNN with class weights, 3D CNN with three slices, and 3D CNN. The graphs were drawn using the validation set results from stratified cross-validation.

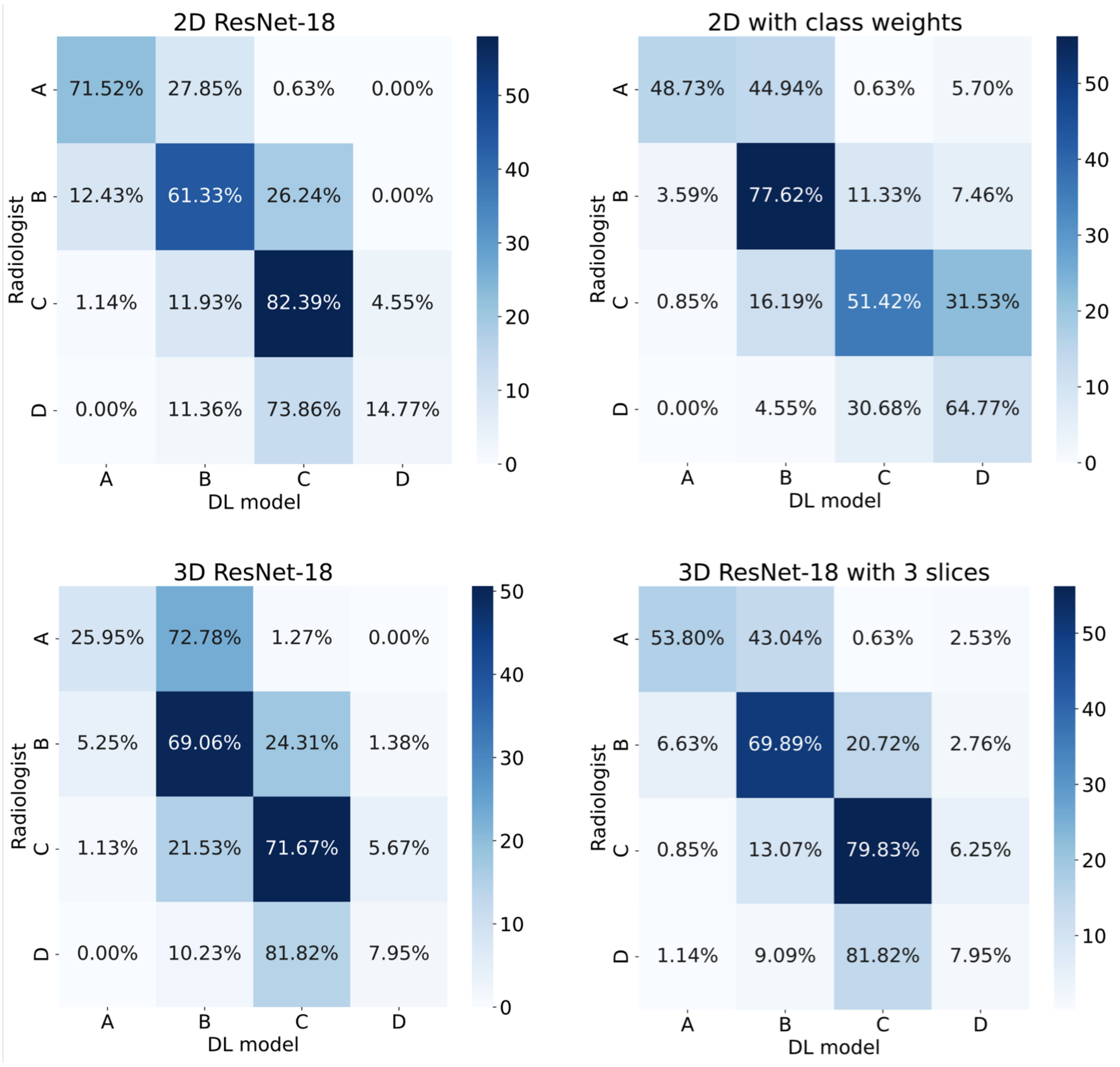
The area under the curve (AUC) values of the ROC curves from Figure 12 are summarised in Table 9. In addition to that, an overall AUC value was calculated for each model. This was done by averaging the individual AUC values from all four classes for a model. Further, the standard deviation for each of these values was also computed. These were computed based on the results from the five folds of the stratified k-fold cross-validation. Each fold gave a result on the validation set, which had an associated ROC curve and AUC value. After the five folds, each model had five separate AUC values, whose average gave the overall AUC value per model. Using Equation 13 on these gave the standard deviation, where N is total folds, x_i is the AUC value from fold i, and $\bar{x}$ is the mean AUC value.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{13}$$

Table 9: The area under the curve values obtained from the ROC curves in Figure 12. Each CNN model gave multiple AUC values for multiple classes. The average of these gave the overall AUC per model. The standard deviation values were included after the ± sign, computed as per Equation 13.

Model	**AUC**				
	A	B	C	D	Overall
2D ResNet-18	0.94±0.02	0.85±0.02	0.84±0.02	0.87±0.03	0.88±0.04
2D ResNet-18 with class weights	0.94±0.02	0.84±0.04	0.82±0.03	0.87±0.03	0.87±0.05
3D ResNet-18 with three slices	0.93±0.01	0.81±0.04	0.83±0.04	0.83±0.07	0.85±0.05
3D ResNet-18	0.72±0.15	0.63±0.10	0.63±0.13	0.70±0.05	0.67±0.04

Additionally, confusion matrices are displayed in Figure 13 averaged over all classes for each model.



2D ResNet-18

Radiologist \ DL model	A	B	C	D
A	71.52%	27.85%	0.63%	0.00%
B	12.43%	61.33%	26.24%	0.00%
C	1.14%	11.93%	82.39%	4.55%
D	0.00%	11.36%	73.86%	14.77%

2D with class weights

Radiologist \ DL model	A	B	C	D
A	48.73%	44.94%	0.63%	5.70%
B	3.59%	77.62%	11.33%	7.46%
C	0.85%	16.19%	51.42%	31.53%
D	0.00%	4.55%	30.68%	64.77%

3D ResNet-18

Radiologist \ DL model	A	B	C	D
A	25.95%	72.78%	1.27%	0.00%
B	5.25%	69.06%	24.31%	1.38%
C	1.13%	21.53%	71.67%	5.67%
D	0.00%	10.23%	81.82%	7.95%

3D ResNet-18 with 3 slices

Radiologist \ DL model	A	B	C	D
A	53.80%	43.04%	0.63%	2.53%
B	6.63%	69.89%	20.72%	2.76%
C	0.85%	13.07%	79.83%	6.25%
D	1.14%	9.09%	81.82%	7.95%

Figure 13: Multiclass confusion matrices for each model using stratified k-fold cross-validation.

6.2 Cross-validation results

As described in the previous section, stratified cross-validation was used for the experiments. The reasons for using cross-validation have been described in section 4.3.8. Mainly, it was to overcome the overfitting problem caused by the scarcity of data and complexity of deep learning models. Further, stratified cross-validation was chosen as the data was highly imbalanced between different classes. The results from the validation sets of five folds were computed for each model. Figure 14 shows the ROC curves for one such model, i.e. 2D CNN. Note, f1, f2, f3, f4, and f5 are the five folds. Alongside these are the AUC values for each fold. These were averaged to give one AUC value per model, which was shown previously in Figure 12.

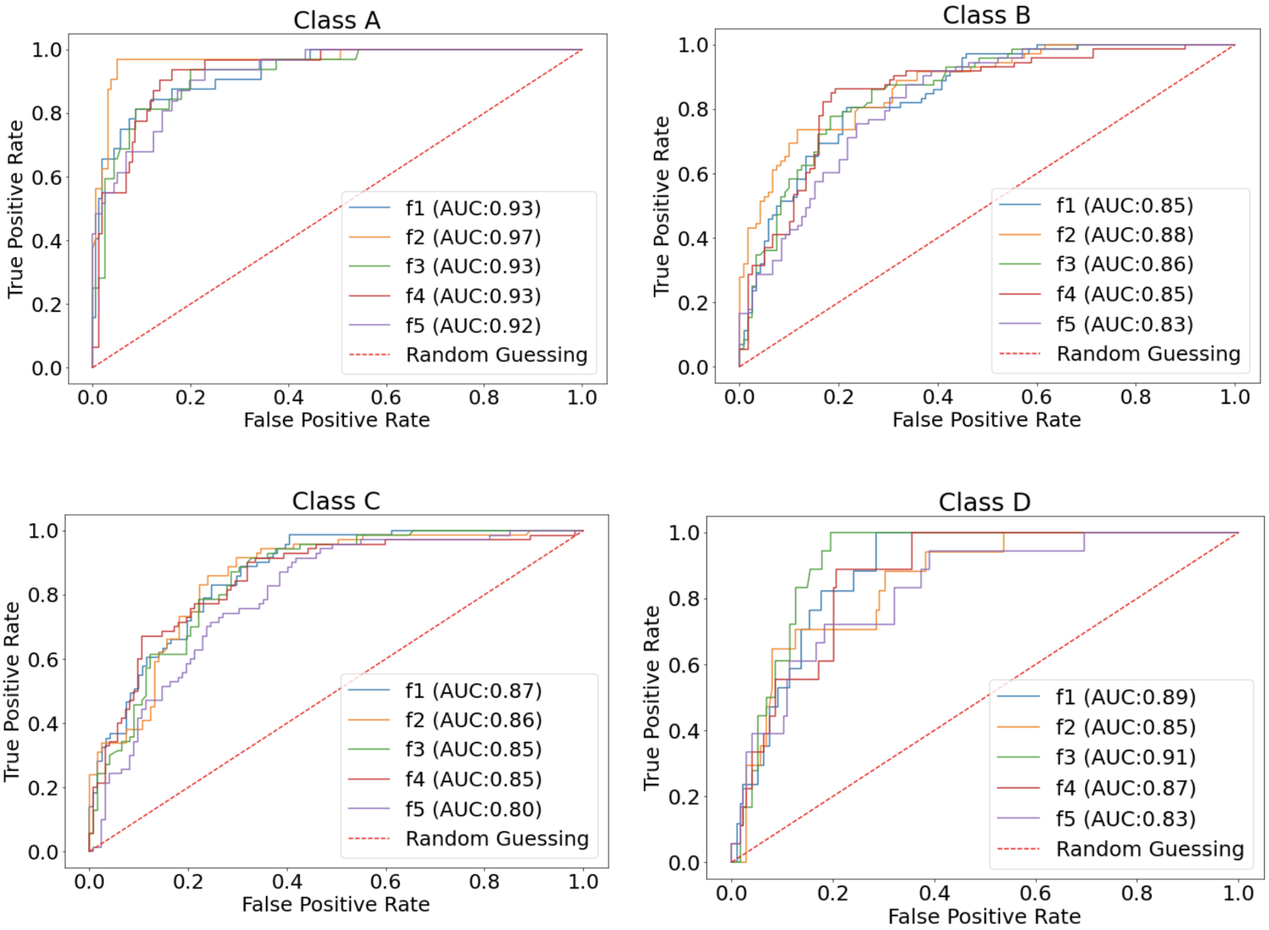


Figure 14: ROC curves for all classes of 2D ResNet-18 for five-fold cross-validation (f denotes fold).

In addition to receiver operating characteristic (ROC) curves, the confusion matrices were also drawn for each model. These described the validation set results from five folds of the stratified cross-validation. The result of one such model, i.e. the 2D CNN is shown in Figure 15. Note the black and white confusion matrices are from the different folds, starting from the first (top left), followed by the second fold, (top right), and so on. The blue confusion matrix is the mean of all the others. It was generated by averaging the TP, TN, FP, and FN values (where, T: true, F: false, P: positive, N: negative) from the five-fold confusion matrices. As explained previously, each class is considered positive at a time, while all the others are considered part of the negative class. These confusion matrices were then combined to give a single confusion matrix per model, as shown in Figure 13.

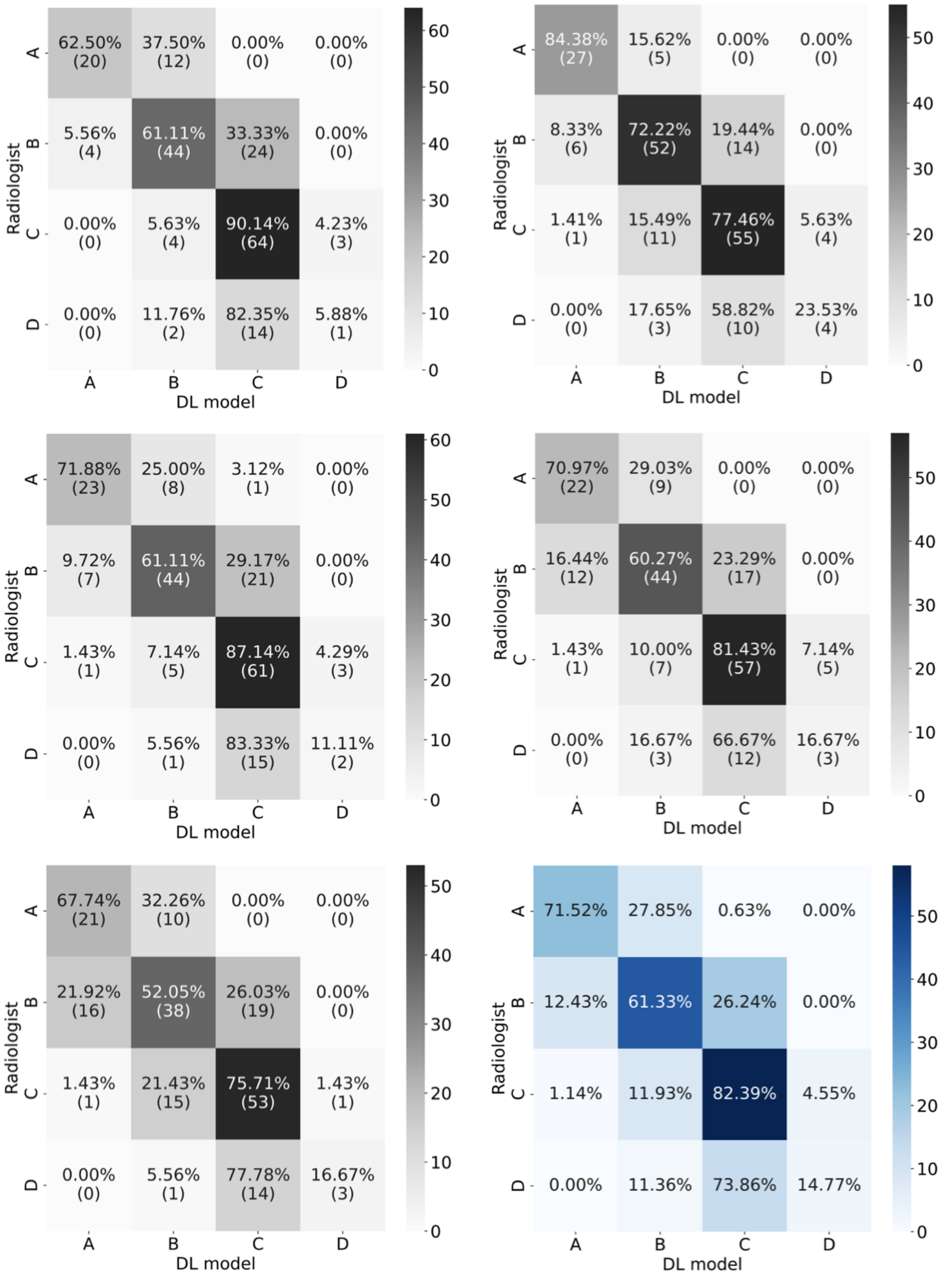


Figure 15: Confusion matrices of 2D ResNet-18 with the transfer learning, showing five-fold results from stratified cross-validation. Their average gave the mean confusion matrix (bottom right).

7 Discussion

As part of this study, a four-class classification problem was investigated on the breast MRI dataset using deep learning. Firstly, a deep learning pipeline was set up to perform preliminary testing on the dataset. For this purpose, the ResNet-18 was selected as the convolutional neural network (CNN). Thereafter, a 2-dimensional (2D) vs 3-dimensional (3D) CNN architecture was compared, along with the use of transfer learning using ImageNet weights. Finally, the use of class weights was studied by applying the technique on one of the 2D CNNs. The inferences observed by looking at the results are summarised in this section.

As per the accuracy, f1 score, precision and recall values as shown in Table 8, the following observations were collected. The 2D CNN with class weights and transfer learning performed the best with *Accuracy* of 0.83 and *f1 score* of 0.61. The 2D CNN with transfer learning performed very similarly, with *Accuracy* of 0.78 and *f1 score* of 0.56. The 3D CNN with three slices performed worse than 2D CNNs with *Accuracy* of 0.75, and *f1 score* of 0.54. The 3D CNN with the whole volume performed the worst with *Accuracy* of 0.70, and *f1 score* of 0.48. Furthermore, the receiver operating characteristic (ROC) curves shown in Figure 12 gave an area under the curve values for each model and class. Looking at these values, as summarised in Table 9, the following observations were collected. The 2D CNN with transfer learning outperformed all with an AUC of 0.88 ± 0.04. It was closely followed by 2D CNN with class weights and transfer learning with an AUC of 0.87 ± 0.05. These were followed by the 3D CNN with three slices with an AUC of 0.85 ± 0.05, which outperformed the 3D CNN with the whole volume with an AUC of 0.67 ± 0.04.

The results for 2D CNNs were as per our assumption. As the inclusion of class weights is supposed to increase the performance of the models, at least on the minority class samples. However, it can also severely affect the classification of the majority class samples. Hence, the overall performance is similar to the 2D CNN without class weights. As the good performance of minority classes (i.e. C and D) is counteracted by the poor performance of the majority classes (i.e. A and B). This was also reflected in the results, as the two models gave similar AUC values with a difference of 0.01. However, a significant difference was observed in the precision, recall, f1-score and accuracy values with approximately 5 points of difference. Nonetheless, we give more weightage to AUC values, as it is a widely preferred metric for classification performance in the medical domain [70]. It also takes into account the performance of each class separately, which can be later combined to give a more robust estimate.

In contrast, the results for 3D CNNs were slightly different from our assumption. It is expected that a more complex model which takes input as 3D instead of a 2D image would give a better performance. Nonetheless, the 3D CNN (with the entire volume of the MRI image as input) gave a worse performance than the 3D CNN with only three middle slices (from the MRI image as the input). Hence, contrary to our belief, the inclusion of all slices did not improve the performance of the model. It decreased the performance significantly in fact. It is also apparent from the ROC graphs in Figure 12 that 3D CNN performed worse than all the others. It is suspected to be an effect of overfitting, as the complexity of the models was high with millions of parameters and the training samples were scarce with only 960 MRI images. The 3D CNN was most affected as it had three times more parameters than the 2D counterparts. Note, unlike the 2D CNNs, the ImageNet weights were not applied to the 3D CNNs. This is assumed to have affected the performance of the two 3D CNNs. However, it still does not account for the poor performance of the 3D CNN compared to 3D3 CNN which had less number of slices. As both were not given the pre-trained ImageNet weights.

The use of class weights is expected to be an encouraged practice in the medical field. As in medical diagnostics, the cost of a false negative or type II error (or miss) is far greater than a false positive or type I error (or false alarm). In some cases, it can be life-threatening e.g. stage 4 cancer. Hence, class weights are preferred to reduce this type of error as much as possible. However, this may lead to more false positives or higher type I error. Hence, a careful trade-off needs to be established between the two errors. In most medical cases this means reducing the number of false negatives at the cost of more false positives. This is because multiple successive screenings can be performed in case of a false positive, which can eventually lead to a true negative later. But for false negatives, the patient might not come again for a screening believing he/she is healthy. This can be lethal for some patients with such false negative results.

For predictive analysis, a benchmark is a previously performed study which could be used as a reference to analyse and compare the results of the current study. However, in our case, no such benchmark was included due to multiple reasons. Firstly, most of the previously done studies were targeted at mammograms and not MRI images, for instance, [5], [11], [17], [18], and [16]. Further, mammographic density can have high variance in images depending on body position, compression levels, and intensity of the X-rays used [9]. Even in our labelled dataset, the mammographic density was not always the same as the MRI density for various scans of the patients. Secondly, the few studies that were performed on the MRI images were also very different from our project. Many of them used quantitative methods along with segmentation such as [9], [13], and [15]. Others used regression instead of classification to estimate breast density such as [14].
Our study aimed to perform classification without segmentation or regression. Mainly because using those techniques was insignificant in answering our research questions such as *“what are the implications of using 2D vs 3D CNN architecture?”*. Another very important reason for not including segmentation was the unavailability of ground truth labels. The time of the radiologists is very limited and often they do not perform segmentation or its labelling as part of their routine. Mainly because most diagnoses are done without segmenting the images. Especially for breast density estimation, which is a fairly new area of research. There is scarce to no availability of segmented patches containing only dense tissues or annotations in the original image. Due to all these reasons, the studies mentioned in the report were not used as the benchmark.

8 Conclusions

In this study, a multiclass classification was performed on the breast MRI data. As part of that, a deep-learning classification pipeline was set up to classify the images into four BI-RADS categories, A, B, C, and D. The ResNet-18 architecture was chosen as the foundation of the different convolutional neural networks. The implications of using a 2-dimensional versus 3-dimensional architecture were studied. Further, the significance of the use of transfer learning (with ImageNet weights) and class weights were gathered. As part of the pipeline, various image processing techniques were also used. These were normalisation, resampling, and augmentation. The ResNet-18 architecture was selected as the basis for our four convolutional neural networks. These were 2D, 2D with class weights, 3D with 3 slices, and 3D CNN. Further, the training of these models was performed using a stratified k-fold cross-validation technique. Mainly to validate the results in the presence of a scarce dataset and class imbalance problem. The output of the models was four floating-point numbers representing the predicted probability of the input image belonging to a particular class.

To compare the results of various ResNet-18 architectures, various metrics were used. These include precision, recall, f1 score, and accuracy. Further, receiver operating characteristic (ROC) curves were used to visualise the classification results. The area under the curve of the ROC graphs was studied to compare the performance of different models. However, since ROC curves are designed for binary classification, they were modified for our multi-class classification problem. This was done by using one class as positive and the remaining three as negative class. Hence, multiple ROC curves per model were obtained for each of the four classes. The average of the area under the curve (AUC) values were obtained for all classes of each model. This was done to easily compare the models with each other. Further note that the results were based on the validation set from five folds of the stratified k-fold cross-validation. Hence, initially, the ROC curve for each fold was obtained separately. These were then combined using the average of the AUC values from all folds. Finally, the ROC curve which had the closest AUC value to the average was chosen as the representative ROC curve for all the folds.

Examining the different graphs and performance metrics, various conclusions were drawn. Specifically, the variability arising from the different architectural choices was investigated. Further, the significance of using transfer learning with or without class weights was assessed. These are further explained in the next subsection along with detailed answers to our research questions. Overall, the results showed a reasonable agreement between predictions from the deep-learning models and the assessment of radiologists. Hence, automatic breast density classification using deep learning can be used as a preliminary or additional diagnostic method in breast density estimation. Note, as it is a supervised learning problem, it can improve greatly with the availability of more labelled data. However, if implemented as an unsupervised or semi-supervised problem, it could help save immense time for radiologists.

8.1 Summary of main contributions

In this section, the main contributions of this study are discussed while giving answers to our research questions. The first research question was targeted towards the applicability of a deep learning convolutional neural network on breast density estimation. The second research question was focused on the implications of architectural choices, such as increasing the dimensionality of the input from 2D to 3D. The third research question was aimed at the relevance of various performance-enhancement techniques such as transfer learning using ImageNet weights and applying class weights (i.e. higher weights to less represented classes like BI-RADS c and d) while training. The answers to each of these research questions are described in detail further in this section.

- *Q1. How does a deep convolutional neural network (CNN) perform on breast MRI data?*

  The classification performance of a deep convolutional neural network was analysed on the given breast magnetic resonance imaging (MRI) data. For this, a ResNet-18 [19] based CNN model was used with slight modifications of input and output layers. Three out of four CNN models gave the area under the curve value of 0.85 or more. In comparison, a random classifier gives only 0.50. Meaning, 85% of the time the classifiers correctly predicted the breast density classes (a,b,c,d). Using these AUC values as the standard, it can be stated that deep convolutional neural networks give moderate to good performance on breast MRI images. Although, a few things should be noted such as the use of augmentation and preprocessing of the images. Further, the number of iterations in the model training was only 50 epochs. This shows that there is still a great amount of improvement that can be seen in the classification performance if training is extended to more epochs say 100, and more real data could be used instead of augmentation.

- *Q2. What are the implications of architectural choices, such as 2- versus 3-dimensional CNN?*

  The significance of using a 2-dimensional versus a 3-dimensional architecture of the CNN model was studied. For this purpose, 2D and 3D ResNet-18 models were used. In all performance metrics listed in the report, it was observed that the 2D model outperformed the 3D model. However, not too much importance should be given to this result, as various factors contributed to the better performance of the 2D model. Firstly, unlike the 3D model, the 2D model used a pretrained ResNet-18 model, in other words, the model was already trained on the huge ImageNet dataset, before being applied to the breast MRI images. Secondly, the 3D model had almost 3 times more parameters than the 2D model, however, the training duration was the same for both, i.e. 50 epochs. Hence, it could be that the 3D model was not being able to learn the useful 3D features in that short duration of time. Additionally, the problem of overfitting might have played a role, where the 3D model memorised the training data instead of learning meaning features from it. However, comparing the performance of the two 3D models, it was observed that the one with only three middle slices of the MRI image as input, performed better than the one using all the MRI image slices as input. Hence, it can be concluded that a 2-dimensional or 2-dimensional-like (i.e. 3D with 3 slices) architecture gives better performance than a 3-dimensional architecture. Therefore, in the end, the architectural choice did contribute significantly to the better classification of breast MRI images.

- *Q3. Does transfer learning and the use of class weights improve the performance of CNN?*

  The importance of using transfer learning was studied with the help of a pretrained ResNet-18. More specifically, the ResNet-18 model was initially trained on the ImageNet dataset [20], which was then used in our study to classify breast density using breast MRI images. It was observed that a 2D model using transfer learning performed better than a 3D model without transfer learning. However, as they used different architectures (2D vs 3D), it cannot be said for certain whether the performance was due to transfer learning or architectural choice. Further, the significance of using class weights was studied using two similar 2D ResNet-18 models. It was observed that the model with class weights gave a slightly better overall performance (in all classes) than the one without class weights. However, the performance increase was drastic in the case of minority classes (i.e. BI-RADS categories c and d). Since in medical diagnostics, the minority classes often signify the presence of a disease. Therefore, having fewer false negatives in these classes is of the highest importance.

8.2 Future work

In this section, the shortcomings of our research setup are exposed along with some suggestions to improve the results of future research targeting similar research questions. First, the quality and quantity of the dataset could be improved. The presence of large noise-free data is of immense significance in any deep-learning study. Further, in medical diagnostics, the challenge of labelling is also seen. Hence, a correctly gathered ground truth from a medical professional is always important. This was the reason segmentation was not used in our study, as there were no ground truth labels related to segmented dense tissue patches. Further, the variability of the dataset could be improved by using datasets from different medical centres. This could help in the generalisation of the problem, and help rule out the presence of noise due to the MRI machines, settings or radiography techniques used at a particular hospital.

Further, in this study, T1-weighted MRI sequences were used as they provided high contrast among fatty and dense tissue. However, the use of other MRI sequences can be explored such as T2-weighted, or TWIST (Time-resolved angiography With Interleaved Stochastic Trajectories). Note, our T1-weighted MRI sequences were non-fat suppressed. However, non-fat-suppressed T1-weighted images show high contrast among fatty and dense tissues because of the bright appearance of the fat [9]. This is absent in fat-suppressed images as they lower the intensity of fatty tissue thereby decreasing the contrast in the image. Moreover, the use of fat-suppressed images is risky as it can alter the intensity of the fatty tissue across the image. This could give rise to a bias field correction issue. However, in some cases fat-suppressed images are preferred such as to decrease the intensity of the liver from the image, to separate it from the breast [9]. The image preprocessing could also be improved such as by using augmentation targeted at the minority classes. This could help with the class imbalance problem, which is prevalent in medical applications. Additionally, MRI-specific preprocessing techniques can be utilised such as slice timing correction (STC), image registration, artefact removal, and bias field correction.

The performance of the 3-dimensional models could be improved using more training iterations such as 100 or more epochs. The fine-tuning of the hyperparameter could also lead to better performance, such as decreasing the learning rate from 0.05. Although, in our case, various learning rates were tested such as 0.001 and 0.01 before choosing 0.05 empirically. However, this was done in the initial phase of the experiments with only one 2D model. It would be helpful to test it on all models separately with the entire pipeline involving k-fold cross-validation. Further, transfer learning and class weights could also be used with 3-dimensional models to improve their performance, and to give a better comparison between 2-dimensional and 3-dimensional models. The choice of ResNet-18 was inspired by previous literature [17]. However, for better classification performance more advanced medically relevant models can be used, e.g. MedicalNet, U-net, or V-net.

The statistical analysis could be improved using tools designed for medical data such as MedCalc [71]. For instance, it could help create better and more intuitive receiver operating characteristic (ROC) curves with standard deviation values embedded in the graph in the shape of a grey area around the lines. Further, other graphical methods can be explored such as the detection error tradeoff (DET) curve. It plots the false positive rate (FPR) versus the false negative rate (FNR).

Bibliography

[1] M. J. Yaffe, J. W. Byng, and N. F. Boyd, "Quantitative image analysis for estimation of breast cancer risk," *Handbook of Medical Imaging, Processing and Analysis*, pp. 323–340, 2000.

[2] R. M. Mann, A. Athanasiou, P. A. Baltzer, J. Camps-Herrero, P. Clauser, E. M. Fallenberg, G. Forrai, M. H. Fuchsjäger, T. H. Helbich, F. Killburn-Toppin, *et al.*, "Breast cancer screening in women with extremely dense breasts recommendations of the european society of breast imaging (eusobi)," *European Radiology*, vol. 32, no. 6, pp. 4036–4045, 2022.

[3] ACS, "Breast cancer - american cancer society." https://www.cancer.org/cancer/breast-cancer. [Online; accessed March 16, 2023].

[4] NHS, "Breast cancer in women - national health service." https://www.nhs.uk/conditions/breast-cancer/, 2019. [Online; accessed March 16, 2023].

[5] A. Hamidinekoo, E. Denton, A. Rampun, K. Honnor, and R. Zwiggelaar, "Deep learning in mammography and breast histology, an overview and future trends," *Medical image analysis*, vol. 47, pp. 45–67, 2018.

[6] NCI, "Breast cancer - national cancer institute." https://www.cancer.gov/types/breast. [Online; accessed March 16, 2023].

[7] C. E. DeSantis, F. Bray, J. Ferlay, J. Lortet-Tieulent, B. O. Anderson, and A. Jemal, "International variation in female breast cancer incidence and mortality rates international variation in female breast cancer rates," *Cancer epidemiology, biomarkers & prevention*, vol. 24, no. 10, pp. 1495–1506, 2015.

[8] N. F. Boyd, G. S. Dite, J. Stone, A. Gunasekara, D. R. English, M. R. McCredie, G. G. Giles, D. Tritchler, A. Chiarelli, M. J. Yaffe, *et al.*, "Heritability of mammographic density, a risk factor for breast cancer," *New England Journal of Medicine*, vol. 347, no. 12, pp. 886–894, 2002.

[9] K. Nie, J.-H. Chen, S. Chan, M.-K. I. Chau, H. J. Yu, S. Bahri, T. Tseng, O. Nalcioglu, and M.-Y. Su, "Development of a quantitative method for analysis of breast density based on three-dimensional breast mri," *Medical physics*, vol. 35, no. 12, pp. 5253–5262, 2008.

[10] "Breast imaging reporting data system — american college of radiology," *American College of Radiology*, 2013.

[11] H.-P. Chan and M. A. Helvie, "Deep learning for mammographic breast density assessment and beyond," 2019.

[12] C. Zhou, H.-P. Chan, N. Petrick, M. A. Helvie, M. M. Goodsitt, B. Sahiner, and L. M. Hadjiiski, "Computerized image analysis: estimation of breast density on mammograms," *Medical physics*, vol. 28, no. 6, pp. 1056–1069, 2001.

[13] A. Gubern-Merida, M. Kallenberg, R. M. Mann, R. Marti, and N. Karssemeijer, "Breast segmentation and density estimation in breast mri: a fully automatic framework," *IEEE journal of biomedical and health informatics*, vol. 19, no. 1, pp. 349–357, 2014.

[14] B. H. van der Velden, M. H. Janse, M. A. Ragusi, C. E. Loo, and K. G. Gilhuijs, “Volumetric breast density estimation on mri using explainable deep learning regression,” *Scientific Reports*, vol. 10, no. 1, p. 18095, 2020.

[15] M. U. Dalmış, G. Litjens, K. Holland, A. Setio, R. Mann, N. Karssemeijer, and A. Gubern-Mérida, “Using deep learning to segment breast and fibroglandular tissue in mri volumes,” *Medical physics*, vol. 44, no. 2, pp. 533–546, 2017.

[16] N. Saffari, H. A. Rashwan, M. Abdel-Nasser, V. Kumar Singh, M. Arenas, E. Mangina, B. Herrera, and D. Puig, “Fully automated breast density segmentation and classification using deep learning,” *Diagnostics*, vol. 10, no. 11, p. 988, 2020.

[17] C. D. Lehman, A. Yala, T. Schuster, B. Dontchos, M. Bahl, K. Swanson, and R. Barzilay, “Mammographic breast density assessment using deep learning: clinical implementation,” *Radiology*, vol. 290, no. 1, pp. 52–58, 2019.

[18] A. A. Mohamed, W. A. Berg, H. Peng, Y. Luo, R. C. Jankowitz, and S. Wu, “A deep learning method for classifying mammographic breast density categories,” *Medical physics*, vol. 45, no. 1, pp. 314–321, 2018.

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[20] “Imagenet large scale visual recognition challenge (ilsvrc).” https://image-net.org/challenges/LSVRC/index.php. [Online; accessed 2-June-2022].

[21] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, “Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012,” *International journal of cancer*, vol. 136, no. 5, pp. E359–E386, 2015.

[22] N. Azamjah, Y. Soltan-Zadeh, and F. Zayeri, “Global trend of breast cancer mortality rate: a 25-year study,” *Asian Pacific journal of cancer prevention: APJCP*, vol. 20, no. 7, p. 2015, 2019.

[23] IHME, “Breast cancer - institute for health metrics and evaluation.” https://www.healthdata.org/, 2015. [Online; accessed March 16, 2023].

[24] WHO, “Breast cancer - world health organisation.” https://www.who.int/news-room/fact-sheets/detail/breast-cancer, 2021. [Online; accessed March 16, 2023].

[25] WCRF, “Breast cancer statistics - world cancer research fund international.” https://www.wcrf.org/cancer-trends/breast-cancer-statistics/, 2021. [Online; accessed March 20, 2023].

[26] CDC, “Basic information about breast cancer - centers for disease control and prevention.” https://www.cdc.gov/cancer/breast/basic_info/, September 2022. [Online; accessed March 16, 2023].

[27] N. F. Boyd, H. Guo, L. J. Martin, L. Sun, J. Stone, E. Fishell, R. A. Jong, G. Hislop, A. Chiarelli, S. Minkin, *et al.*, “Mammographic density and the risk and detection of breast cancer,” *New England journal of medicine*, vol. 356, no. 3, pp. 227–236, 2007.

[28] NIH, "Breast cancer: What role does breast density play? - national center for biotechnology information." https://www.ncbi.nlm.nih.gov/books/NBK447118/, March 2017. [Online; accessed March 16, 2023].

[29] P. E. Freer, "Mammographic breast density: impact on breast cancer risk and implications for screening," *Radiographics*, vol. 35, no. 2, pp. 302–315, 2015.

[30] J. O. Wanders, K. Holland, W. B. Veldhuis, R. M. Mann, R. M. Pijnappel, P. H. Peeters, C. H. van Gils, and N. Karssemeijer, "Volumetric breast density affects performance of digital screening mammography," *Breast cancer research and treatment*, vol. 162, pp. 95–103, 2017.

[31] Y. Xu, Y. Wang, J. Yuan, Q. Cheng, X. Wang, and P. L. Carson, "Medical breast ultrasound image segmentation by machine learning," *Ultrasonics*, vol. 91, pp. 1–9, 2019.

[32] B. Liu, H.-D. Cheng, J. Huang, J. Tian, X. Tang, and J. Liu, "Probability density difference-based active contour for ultrasound image segmentation," *Pattern Recognition*, vol. 43, no. 6, pp. 2028–2042, 2010.

[33] C. K. Kuhl, S. Schrading, C. C. Leutner, N. Morakkabati-Spitz, E. Wardelmann, R. Fimmers, W. Kuhn, and H. H. Schild, "Mammography, breast ultrasound, and magnetic resonance imaging for surveillance of women at high familial risk for breast cancer," *Journal of clinical oncology*, vol. 23, no. 33, pp. 8469–8476, 2005.

[34] E. Paci, M. Broeders, S. Hofvind, D. Puliti, S. W. Duffy, and E. W. Group, "European breast cancer service screening outcomes: a first balance sheet of the benefits and harms," *Cancer epidemiology, biomarkers & prevention*, vol. 23, no. 7, pp. 1159–1163, 2014.

[35] S. Saadatmand, H. A. Geuzinge, E. J. Rutgers, R. M. Mann, D. B. d. R. van Zuidewijn, H. M. Zonderland, R. A. Tollenaar, M. B. Lobbes, M. G. Ausems, M. van't Riet, *et al.*, "Mri versus mammography for breast cancer screening in women with familial risk (famrisc): a multicentre, randomised, controlled trial," *The Lancet Oncology*, vol. 20, no. 8, pp. 1136–1147, 2019.

[36] D. Chicco, "Ten quick tips for machine learning in computational biology," *BioData mining*, vol. 10, no. 1, p. 35, 2017.

[37] R. C. Gonzales and P. Wintz, *Digital image processing*. Addison-Wesley Longman Publishing Co., Inc., 1987.

[38] "About dicom: Overview." [Online; accessed March 1, 2023].

[39] "Nifti: Neuroimaging informatics technology initiative." https://nifti.nimh.nih.gov/. [Online; accessed March 1, 2023].

[40] S. Saponara and A. Elhanashi, "Impact of image resizing on deep learning detectors for training time and model performance," in *Applications in Electronics Pervading Industry, Environment and Society: APPLEPIES 2021*, pp. 10–17, Springer, 2022.

[41] D. Dumitrescu and C.-A. Boiangiu, "A study of image upsampling and downsampling filters," *Computers*, vol. 8, no. 2, 2019.

[42] M. Hashemi, "Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation," *Journal of Big Data*, vol. 6, no. 1, pp. 1–13, 2019.

[43] “Welcome to opencv documentation!.” https://docs.opencv.org/2.4/index.html#. [Online; accessed March 1, 2023].

[44] E. H. Meijering, “Spline interpolation in medical imaging: comparison with other convolution-based approaches,” in *2000 10th European Signal Processing Conference*, pp. 1–8, IEEE, 2000.

[45] “Scipy.ndimage.zoom.” https://docs.scipy.org/doc/scipy/reference/generated/scipy.ndimage.zoom.html#scipy.ndimage.zoom. [Online; accessed Feb 15, 2023].

[46] A. B. Jung, “imgaug.” https://github.com/aleju/imgaug, 2020. [Online; accessed March 15, 2023].

[47] S. H. Javaheri, M. M. Sepehri, and B. Teimourpour, “Chapter 6 - response modeling in direct marketing: A data mining-based approach for target selection,” in *Data Mining Applications with R* (Y. Zhao and Y. Cen, eds.), pp. 153–180, Boston: Academic Press, 2014.

[48] N. V. Chawla, *Data Mining for Imbalanced Datasets: An Overview*, pp. 875–886. Boston, MA: Springer US, 2010.

[49] C. X. Ling and C. Li, “Data mining for direct marketing: Problems and solutions.,” in *Kdd*, vol. 98, pp. 73–79, 1998.

[50] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[51] R. van den Goorbergh, M. van Smeden, D. Timmerman, and B. Van Calster, “The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression,” *Journal of the American Medical Informatics Association*, vol. 29, no. 9, pp. 1525–1534, 2022.

[52] Y. Lecun, K. Kavukcuoglu, and C. Farabet, “Convolutional networks and applications in vision,” pp. 253–256, 05 2010.

[53] K. Gurney, *An introduction to neural networks*. CRC press, 1997.

[54] D. C. Lay, S. R. Lay, and J. McDonald, *Linear algebra and its applications*. Pearson Education, 2016.

[55] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[56] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[57] K. Janocha and W. M. Czarnecki, “On loss functions for deep neural networks in classification,” *arXiv preprint arXiv:1702.05659*, 2017.

[58] 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*, vol. 1, no. 4, pp. 541–551, 1989.

[59] V. Dumoulin and F. Visin, “A guide to convolution arithmetic for deep learning,” *arXiv preprint arXiv:1603.07285*, 2016.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[61] Y. Chandola, J. Virmani, H. Bhadauria, and P. Kumar, "4.4.2 directed acyclic graph end-to-end pre-trained cnn model: Resnet18 - deep learning for chest radiographs," *Computer Aided Classification, Academic Press*, 2021.

[62] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the royal statistical society: Series B (Methodological)*, vol. 36, no. 2, pp. 111–133, 1974.

[63] F. Ramzan, M. U. G. Khan, A. Rehmat, S. Iqbal, T. Saba, A. Rehman, and Z. Mehmood, "A deep learning approach for automated diagnosis and multi-class classification of alzheimer's disease stages using resting-state fmri and residual neural networks," *Journal of medical systems*, vol. 44, pp. 1–16, 2020.

[64] "Python documentation," March 2023. [Online; accessed March 16, 2023].

[65] F. Chollet *et al.*, "Keras." https://keras.io, 2015. [Online; accessed June 16, 2022].

[66] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems." https://www.tensorflow.org/, 2015. [Online; accessed June 16, 2022].

[67] R. Solovyev, A. A. Kalinin, and T. Gabruseva, "3d convolutional neural networks for stalled brain capillary detection," *Computers in Biology and Medicine*, vol. 141, p. 105089, 2022.

[68] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote sensing of Environment*, vol. 62, no. 1, pp. 77–89, 1997.

[69] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[70] K. Hajian-Tilaki, "Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation," *Caspian journal of internal medicine*, vol. 4, no. 2, p. 627, 2013.

[71] M. S. Ltd, "Medcalc, easy-to-use statistical software." https://www.medcalc.org/, March 2023. [Online; accessed March 15, 2023].

Appendices

A 2D ResNet-18

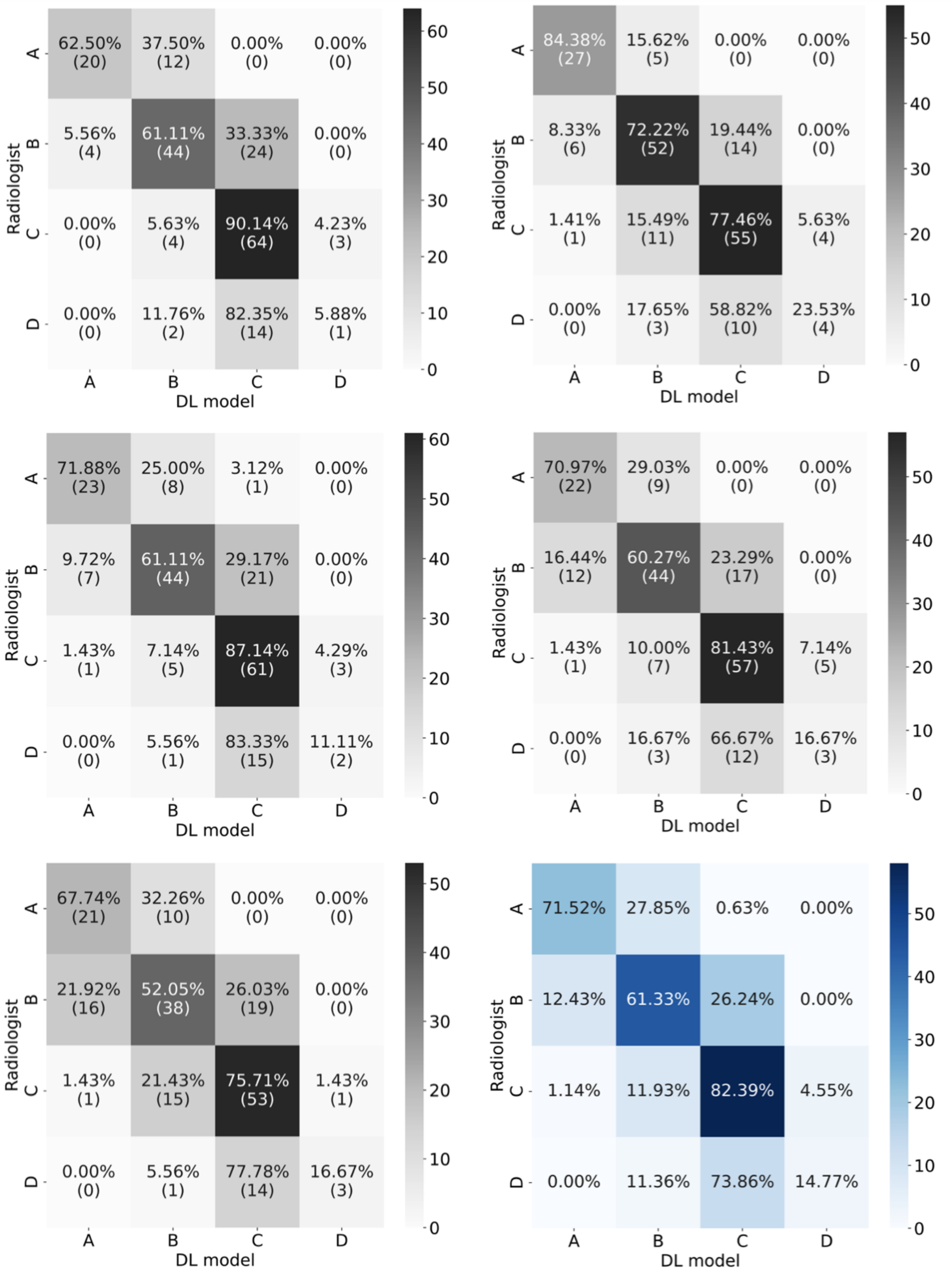


Figure 16: Confusion matrices of 2D ResNet-18 for five-fold cross-validation and mean confusion matrix (bottom right).

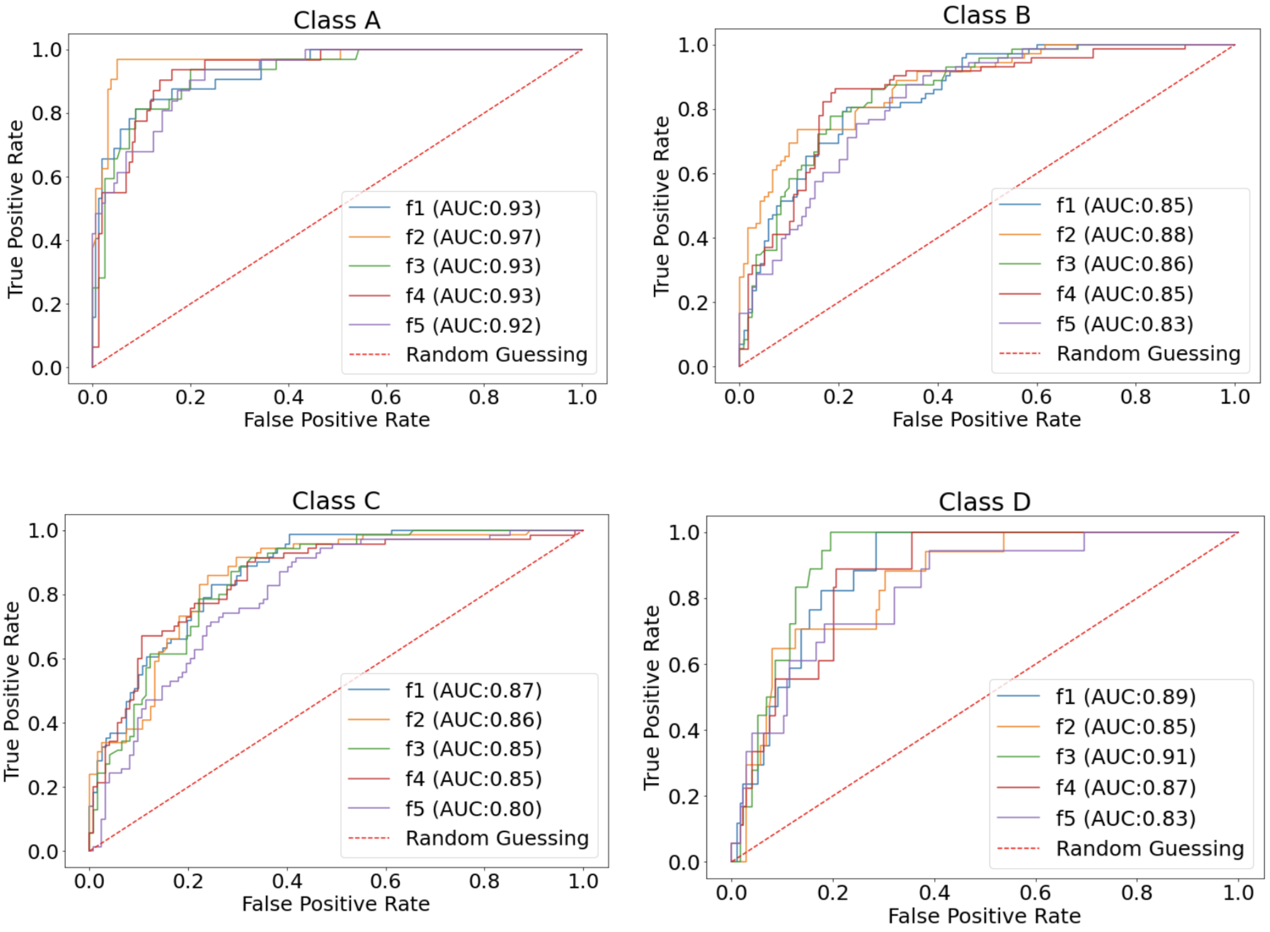


Figure 17: ROC curves for all classes of 2D ResNet-18 for five-fold cross-validation (f denotes fold).

B 2D ResNet-18 with class weights

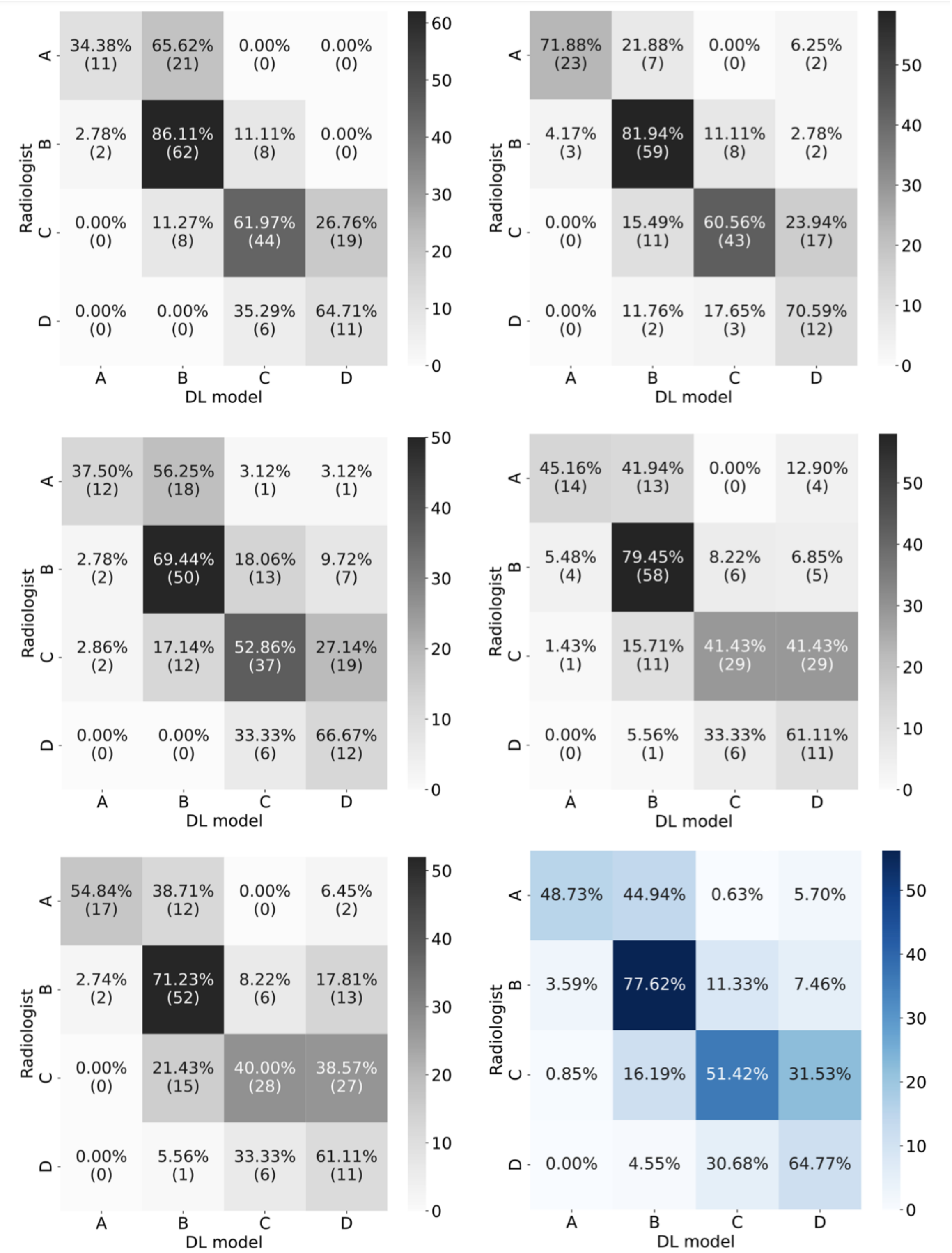


Figure 18: Confusion matrices of 2D ResNet-18 with class weights for five-fold cross-validation and mean confusion matrix (bottom right).

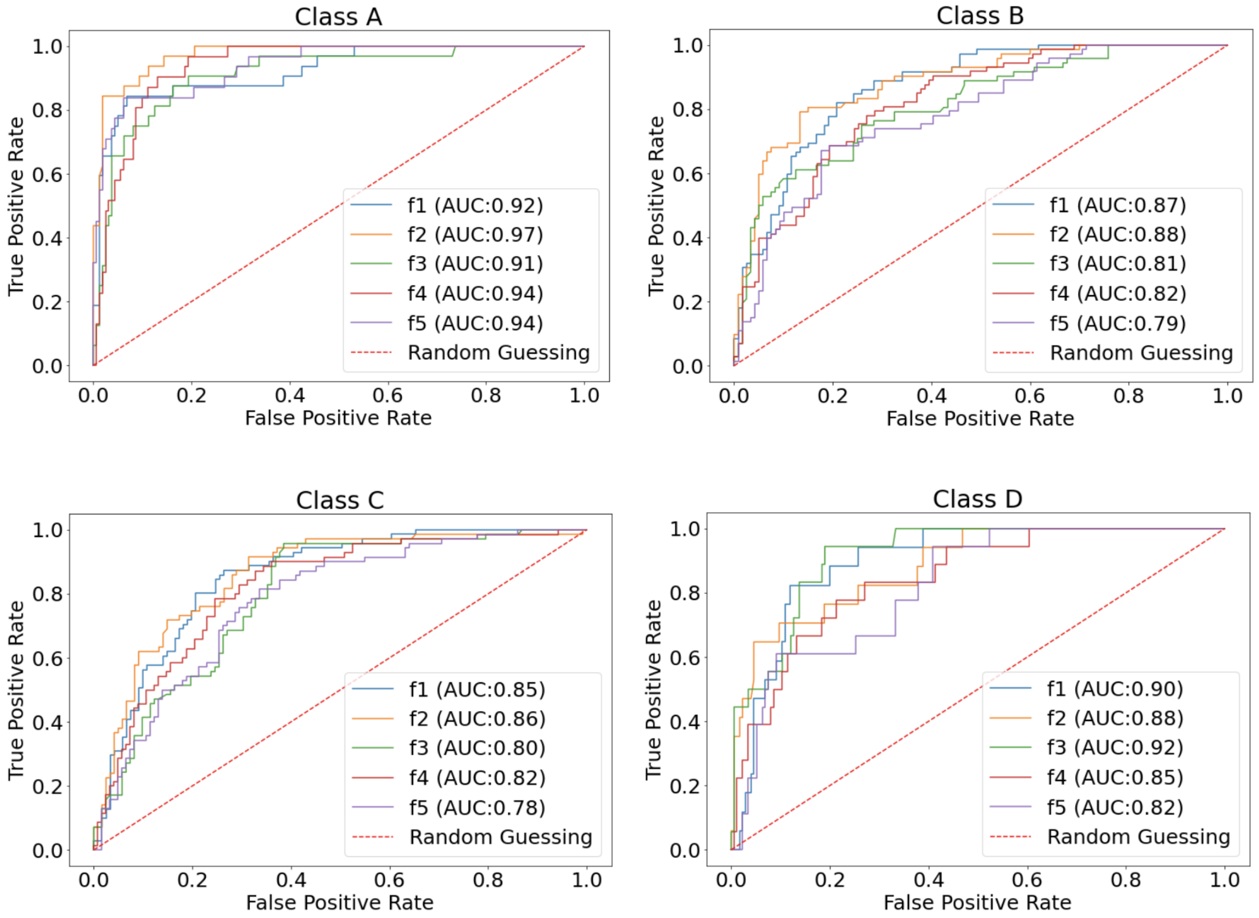


Figure 19: ROC curves for all classes of 2D ResNet-18 with class weights for five-fold cross-validation (f denotes fold).

C 3D ResNet-18 with three slices

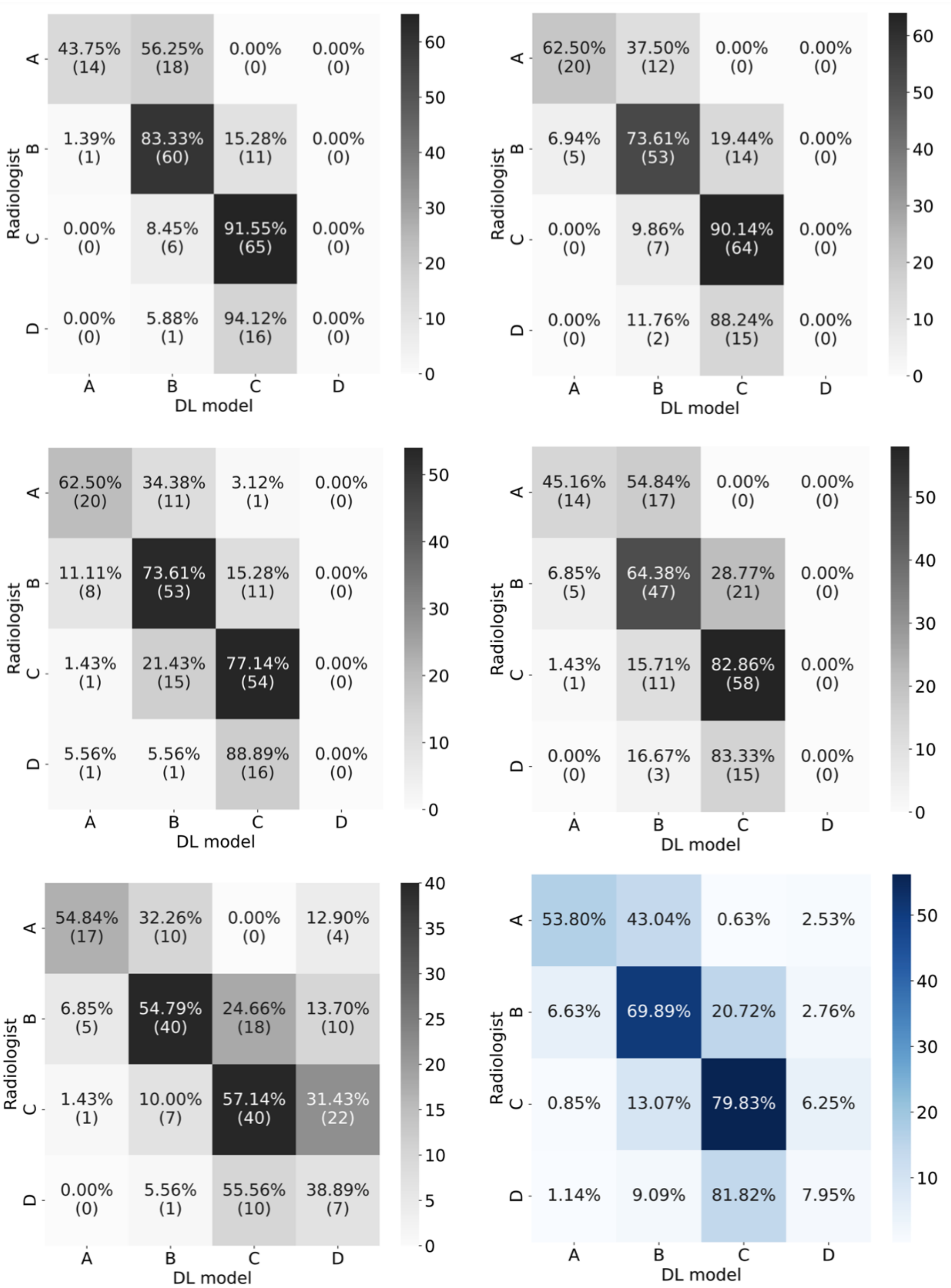


Figure 20: Confusion matrices of 3D ResNet-18 with 3 slices for five-fold cross-validation and mean confusion matrix (bottom right).

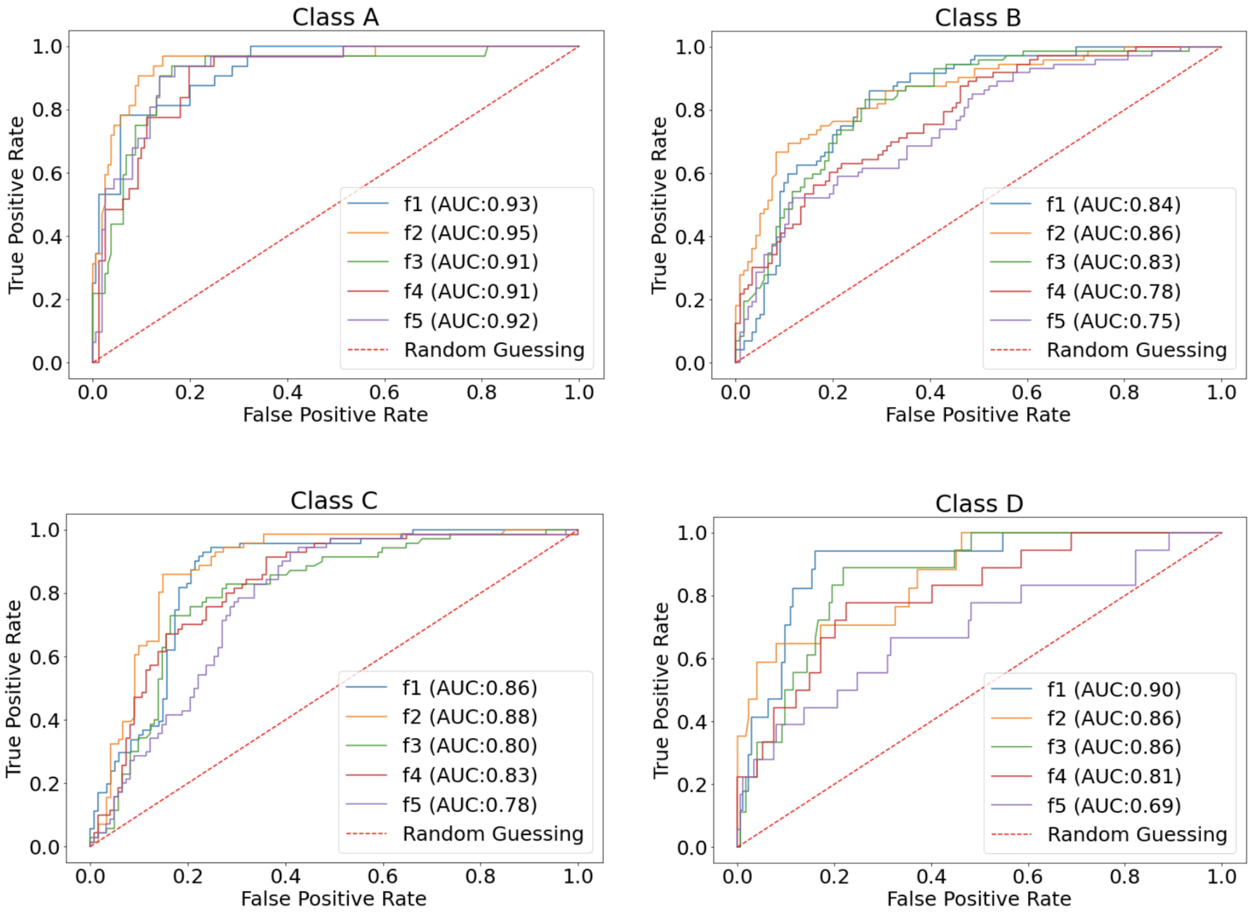


Figure 21: ROC curves for all classes of 3D ResNet-18 with three slices for five-fold cross-validation (f denotes fold).

D 3D ResNet-18 with the whole volume

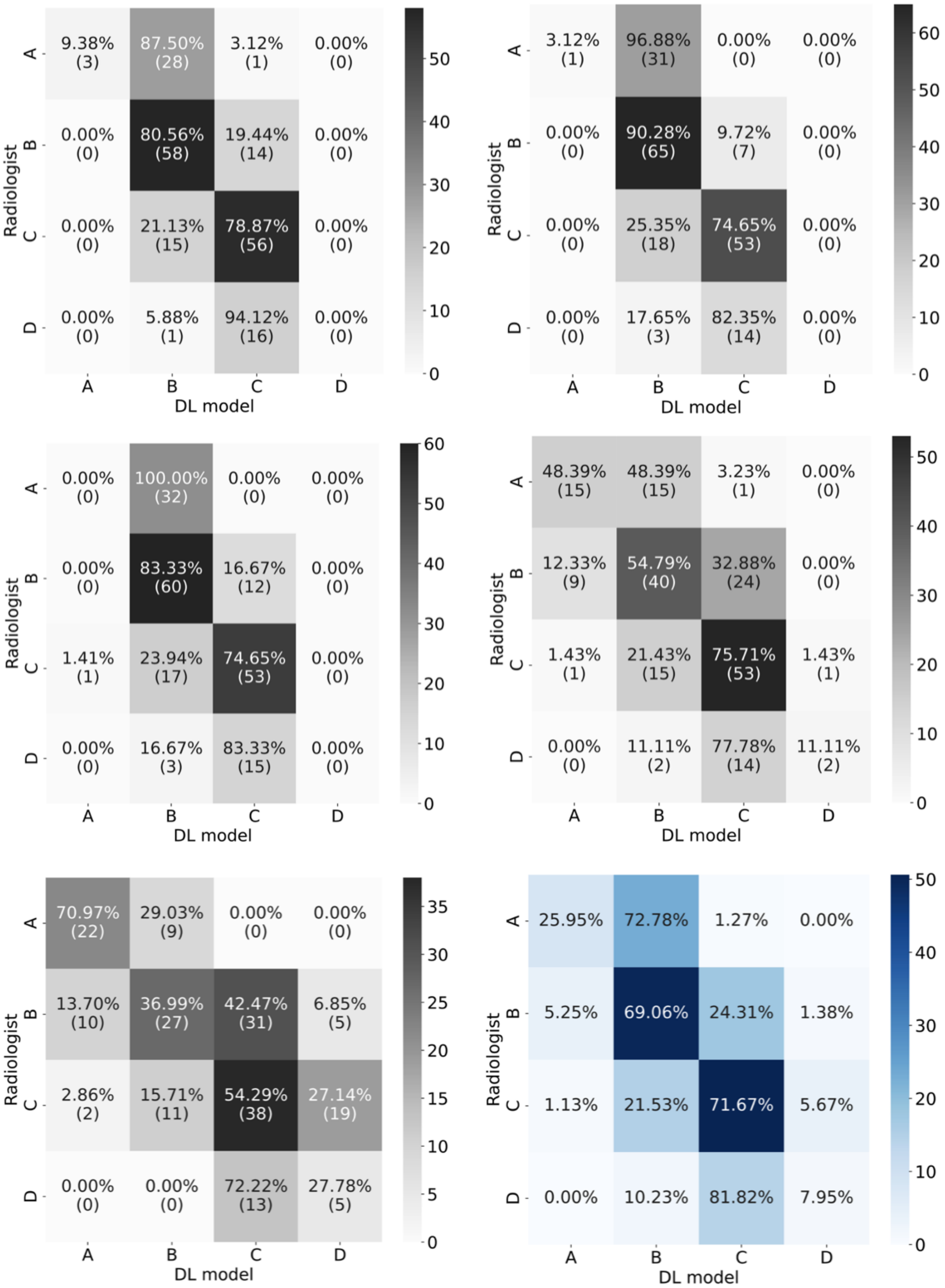


Figure 22: Confusion matrices of 3D ResNet-18 for five-fold cross-validation and mean confusion matrix (bottom right).

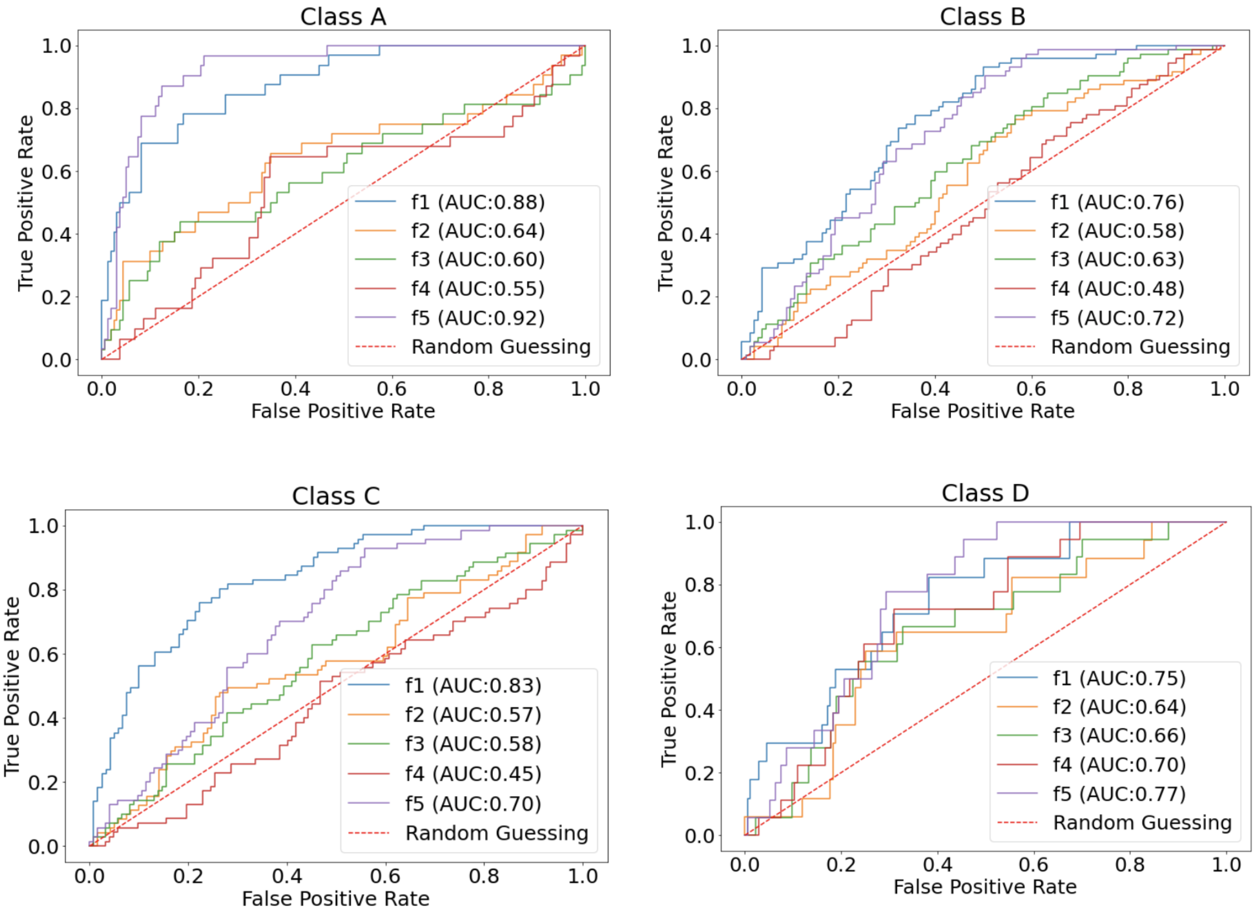


Figure 23: ROC curves for all classes of 3D ResNet-18 for five-fold cross-validation (f denotes fold).