DEVELOPMENT AND OPTIMIZATION OF AI MODELS BASED ON MEDICAL IMAGES

by

NIKOLAOS OIKONOMOU

BSc. Technological Educational Institute of Crete, 2023

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

SCHOOL OF ENGINEERING

HELLENIC MEDITERRANEAN UNIVERSITY

2025

Approved by:

Major Professor
Kostas Marias

Abstract

Prostate cancer (PCa) and therapy-induced cardiac toxicity represent two critical challenges in precision oncology: the former as one of the most frequent male cancers worldwide and the latter as severe side effect of breast cancer treatment that significantly impacts patient outcomes. Artificial Intelligence (AI) offers promising solutions to both problems, however its adoption is limited by heterogeneous imaging data and the lack of dedicated echocardiographic-based models for cardiotoxicity prediction.

This master's thesis addresses these gaps through two complementary studies using datasets from the EU-funded ProCAncer-i and CARDIOCARE projects. For PCa management, bi-parametric MRI scans (T2W, ADC, DWI) were employed to train four 3D deep learning models under different harmonization strategies, including z-score, mean and histogram normalization and the combination of z-score and N4 bias field correction. For the cardiotoxicity risk prediction, 3D echocardiographic videos acquired from breast cancer patients were employed, using 3D ResNet18 and I3D architectures to predict cardiotoxicity occurrence within one year of the baseline examination.

The results show the significance of harmonization techniques in enhancing AI models generalization ability on heterogeneous MRI data. Specifically, in PCa management, ResNet18 trained on histogram-normalized 24x192x192 MR images, achieved an AUC of 71.28%, compared to 63.52% with raw data. In cardiotoxicity risk prediction, ResNet18 outperformed I3D and EchoNet pre-trained models, achieving a mean AUC of 64.52% using the A4C view at low resolution data. Taken together, these findings highlight the value of harmonization for PCa AI models and demonstrate the feasibility of video-based deep learning models as a supportive tool for cardiotoxicity monitoring.

Keywords: prostate cancer biological aggressiveness classification, breast cancer therapy-induced cardiotoxicity risk prediction, 3D deep learning architectures, ProstateNET, CARDIOCARE, 3D echocardiographic videos, 3D prostate cancer MRI scans

Table of Contents

Abstract ii
Table of Contents iii
List of Figures v
List of Tables vii
Acknowledgements viii
Dedication ix
1. Introduction 1
2. Background 2
2.1 Artificial Intelligence in Medical Imaging 2
2.1.1 Deep Learning Architectures 5
2.1.1.1 Convolutional Neural Networks 5
2.1.1.2 Transformers Neural Networks 11
2.1.2 AI challenges in medical imaging 13
2.1.3 Pre-processing Techniques 15
2.1.4 Harmonization Techniques 18
2.1.5 Evaluation Metrics 21
2.2 Prostate Cancer Management 24
2.2.1 Introduction 24
2.2.2 Current advances in AI applications for PCa management 25
2.2.3 Notable Datasets 27
2.3 Therapy-Induced Cardiac Toxicity 29
2.3.1 Introduction 29
2.3.2 Cardiotoxicity risk assessment 31
2.3.3 Notable datasets 33
3. Harmonization techniques and their impact on clinically significant PCa classification 35
3.1 Introduction 35
3.2 ProstateNET Imaging Archive 37
3.3 Pre-processing 40

3.4 Implemented Harmonization Techniques 41
3.5 Implementation details 44
3.6 Results 48
3.7 Discussion 54
4. Cardiotoxicity risk prediction 57
4.1 Introduction 57
4.2 CARDIOCARE dataset 59
4.3 Pre-processing 61
4.4 Implementation details 63
4.5 Results 66
4.6 Discussion 68
5. Conclusion 72
5.1 Study Limitations 73
5.2 Future Work 73
References 75

List of Figures

Figure 1. The relationship between CV and AI 2
Figure 2. Machine Learning workflow 3
Figure 3. ANNs inspired by biological neurons of human brain 4
Figure 4. Supervised and Unsupervised learning 5
Figure 5. Simple CNN architecture for medical image classification 6
Figure 6. A visual representation of a convolutional layer 7
Figure 7. A visual representation of a max-pooling layer operation 7
Figure 8. A pooled feature map being flattened and given as input to fully connected layer 8
Figure 9. Architectural comparison of ResNet variants (18, 34, 50, 101, and 152 layers) [5] 9
Figure 10. EfficientNet - B0 architecture 9
Figure 11. VGG-11 architecture 10
Figure 12. I3D Architectures: (a) Single-stream RGB Model (our implementation), (b) Full Two-stream Model (RGB + Optical Flow) 11
Figure 13. Transformer structure 12
Figure 14. ViT framework 13
Figure 15. Examples of geometric data augmentation on chest X-ray image 16
Figure 16. Example of image cropping on chest X-ray 17
Figure 17. Common pipeline used in multi-center studies 19
Figure 18. Illustration of the piece-wise linear mapping 21
Figure 19. Four main categories of AI tasks in clinical cardiac radiology [62] 31
Figure 20. Overview of the proposed pipeline 36
Figure 21. Dataset Class Distribution 38
Figure 22. Distribution of PCa Patients by MRI Scanner Strength and Clinical Significance 39
Figure 23. Number of cases per manufacturer 39
Figure 24. MRI sequences at three volume sizes (slice 16). (a) raw data 32x224x224. (b) center-crop data 24x192x192. (c) center-crop data 24x128x128 40
Figure 25. All MRI sequences - Raw data 41
Figure 26. All MRI sequences - After z-score normalization 42
Figure 27. All MRI sequences - After histogram normalization 42

Figure 28. All MRI sequences - After mean normalization 43
Figure 29. All MRI sequences - After z-score & N4 bias field correction 43
Figure 30. Study population distribution and dataset stratification 44
Figure 31. The 3-channel architecture 45
Figure 32. Overview of the proposed pipeline in cardiotoxicity risk prediction 58
Figure 33. Cardiotoxicity dataset class distribution 60
Figure 34. Ultrasounds overall quality 61
Figure 35. A 28-frame video sequence padded with zero-valued frames to reach the most frequent number of frames (34) 62
Figure 36. A4C view study population distribution and dataset stratification 63
Figure 37. A2C view study population distribution and dataset stratification 63
Figure 38. a) Model using only ultrasound video data, b) Model using both ultrasound video data and tabular clinical measurements (LV_GLS and LVEF) 64

List of Tables

Table 1. Confusion Matrix 22
Table 2. ProstateNET Scanner Models by Manufacturer 37
Table 3. Classification results - 24x128x128 data 49
Table 4. Classification results - 24x192x192 data 51
Table 5. Classification results - 32x224x224 data 53
Table 6. A2C view - Video sequences only data 66
Table 7. A2C view - Video sequences + Tabular Data 66
Table 8. A4C view - Video sequences only data 67
Table 9. A4C view - Video sequences + Tabular Data 67

Acknowledgements

I would like to extend my sincere thanks to my supervisor professor Dr. Kostas Marias, who inspired me to explore the medical imaging field during my master's studies and gave me the chance to work on such interesting and scientifically critical research challenges.

I would also like to express my profound appreciation to my professor Dr. Manolis Tsiknakis for his guidance, support and for giving me the opportunity to be a member of the Biomedical Informatics & eHealth Laboratory (BMI Lab) at Hellenic Mediterranean University (HMU), where I have had the chance to get involved in research over the past years.

I am sincerely indebted to Dr. Grigorios Kalliatakis, whose precious contributions have been essential to the completion of this thesis. Finally, I am especially grateful to all the members of BMI Lab as well as to all those who have supported me throughout my master's studies.

Dedication

Dedicated to all who believed in me and supported me throughout my studies, and to all researchers working to address the challenges explored in this thesis. Their commitment and passion inspired every step of this work.

1. Introduction

Artificial Intelligence (AI) has become one of the most prominent research areas in medical imaging, demonstrating remarkable progress in domains such as computer vision, natural language processing and medicine. Within medical imaging, AI algorithms have been extensively applied in various modalities including MRI, CT and ultrasound to automatically recognize complex patterns, with established applications in oncology, pathology and radiology. AI is increasingly regarded as a valuable clinical tool, supporting disease diagnosis, image segmentation, outcome prediction, and disease classification. However, a notable gap in literature remains, concerning the impact of pre-processing and harmonization techniques on improving AI models performance. Moreover, no automated AI approaches have yet been developed to predict therapy-induced cardiotoxicity in breast cancer patients using echocardiographic video sequences.

The objectives of this master's thesis are two-fold: 1) to design and develop pre-processing techniques for harmonizing heterogeneous prostate cancer (PCa) imaging data to improve AI model performance and 2) to develop an ultrasound-based AI model that uses entire echocardiographic video sequences to predict cardiotoxicity occurrence within 1 year of the baseline examination. To this end, a range of techniques, experimental scenarios and deep learning architectures were investigated and systematically compared across both tasks.

This thesis is structured as follows: Section 2 presents the background on AI in medical imaging, prostate cancer and cardiotoxicity along with a review of related work for both tasks. Also, it describes the pre-processing and harmonization techniques employed as well as the evaluation metrics. Section 3 details the implementation of the first task, prostate cancer management using AI, including the data description, methodology and results. Section 4 refers to the second task, cardiotoxicity risk prediction, presenting the dataset used, methodological approach and experimental findings. Finally, Section 5 summarizes the conclusions and achievements of this work, discusses its limitations and outlines directions for future work.

2. Background

This section sets the necessary background for this thesis. Firstly, it presents the role and challenges of artificial intelligence and describes the main deep learning architectures, followed by the pre-preprocessing methods, harmonization techniques and the evaluation metrics employed in this study. Continuing, it outlines the background of prostate cancer and cancer therapy-induced cardiotoxicity, describing the current advances, commonly used datasets and challenges found in literature.

2.1 Artificial Intelligence in Medical Imaging

Artificial Intelligence (AI) has a strong presence in many scientific journals, especially in medical imaging during last years. 'AI' as term, coined by John Macarthy in 1956 and refers to computer algorithm's ability to mimic characteristic of human intelligence for learning or problem solving. Recent advancements in computational power and data availability have made AI successful. Modern medicine has been revolutionized by the development of Computer Vision (CV) and AI, which enables better appreciation of anatomy, as well as accurate detection and treatment of diseases at an early stage. Machine learning-based applications including disease diagnosis, image segmentation, outcome prediction have been developed, providing support in clinical decision-making [1][2]. AI is categorized into two subfields, Machine Learning (ML) and Deep Learning (DL). The figure below illustrates the relationship between CV and AI.

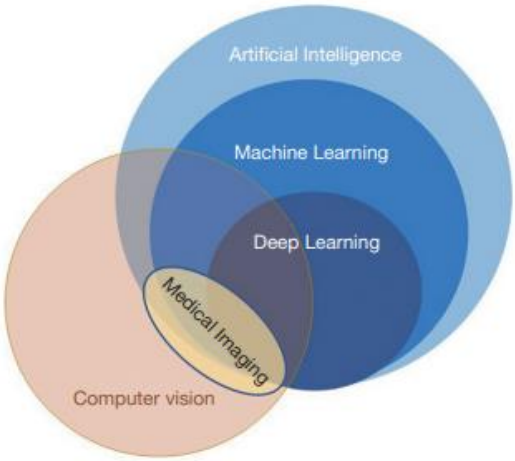


Figure 1. The relationship between CV and AI

ML is part of AI that algorithms can automatically learn from data and observations, however requiring humans to define and extract features from raw data[2][3]. Support Vector Machines (SVM), Random Forest (RF), K-Means, k-nearest neighbors (k-NN), Linear & Logistic Regression are among the most widely used ML algorithms. On the other hand, DL is a subfield of ML which automatically can extract meaningful features from data using artificial neural networks (ANNs) [1][2]. ANNs are computational models inspired by the function of biological nervous systems like the human brain. They are composed of numerous interconnected computational nodes known as neurons that are closely linked and processes information in parallel. Through this structure, ANNs learn from input data and then iteratively change the output in order to increase performance [4].

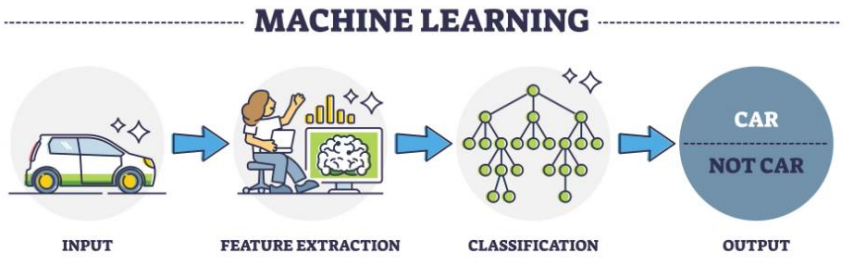


Figure 2. Machine Learning workflow [1]

Compared to ML algorithms, DL networks integrate all the necessary steps for predictive modeling including feature extraction and learning, into an end-to-end network. A DL model consists of a stack of processing layers. The first layer is called *input* and the last layer called *output* while all the layers between them are called *hidden layers*. On account of their multi-layered structure, are able to learn different patterns of the input data. Nevertheless, DL models are susceptible to overfitting especially when the size of training data is relatively small. Due to their ability to analyze complex patterns in large datasets, DL has been widely used in medical imaging and radiation oncology [3]. DL models have demonstrated high diagnostic performance in applications such as skin cancer detection, organ segmentation and mammogram classification,

[1] Source: https://www.turing.com/kb/ultimate-battle-between-deep-learning-and-machine-learning accessed 10/2/2025

outperforming even medical imaging specialists. Inspired by human learning, ML is divided into two major categories of learning processes, supervised and unsupervised learning [1].

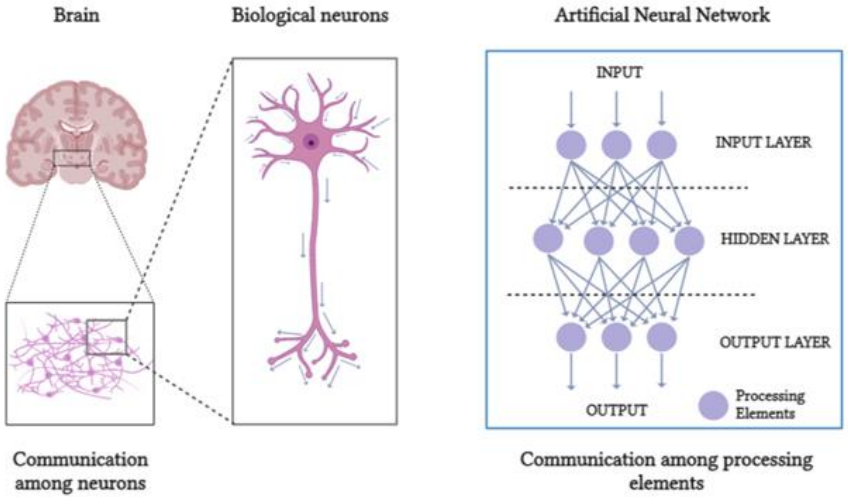


Figure 3. ANNs inspired by biological neurons of human brain [2]

In supervised learning models are provided with training data, where each data sample has been labelled for a specific class or category. The learning algorithm aims to find the relationship that links each input from the training set to its corresponding output label. For instance, in medicine, input data can be in the form of clinical records or medical images while the output label can be in terms of disease diagnosis, patient condition (e.g. disease progression at a certain follow-up), or treatment outcomes (e.g. recurrence or survival rates). After a model is trained to identify these patterns, it is used to new, unlabeled data to classify it into one of the classes that were pre-established during training.

Unsupervised learning, in contrast to supervised learning data used for training, do not belong to any pre-defined class or category. Continuing, the learning algorithm is fed with the training set of data and searches for hidden patterns that possibly can divide data into different groups based on their common characteristics. Once the training phase is completed, the algorithm is ready to classify new unseen data into one of the defined groups that were formed during testing phase [3].

[2] Source: image https://www.researchgate.net/figure/llustrative-image-of-artificial-neural-networks-with-their-working-mechanism-as-like_fig13_380136894

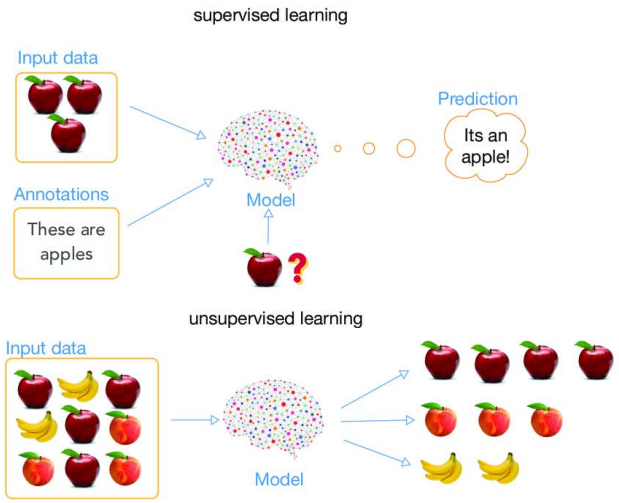


Figure 4. Supervised and Unsupervised learning [3]

2.1.1 Deep Learning Architectures

2.1.1.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) is the most popular DL architecture for medical image processing applications, inspired by human visual system [1]. CNNs consists of stack layers of convolutions (convolutional layers), pooling layers and fully connected layers to the output. Below figure illustrates a simple CNN architecture for image classification.

[3] Source: https://www.researchgate.net/figure/Supervised-learning-and-unsupervised-learning-Supervised-learning-uses-annotation_fig1_329533120

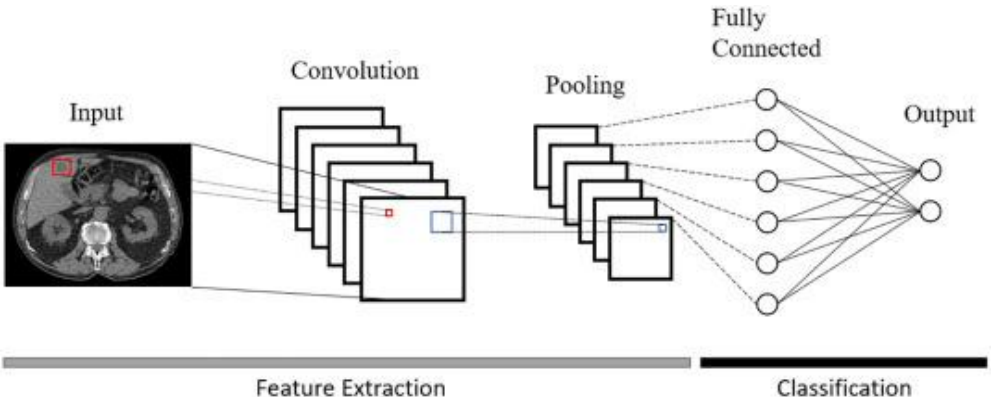


Figure 5. Simple CNN architecture for medical image classification [4]

As seen in Figure 5, network takes an image as input which is processed through convolutional layer to extract feature maps. Continuing, through pooling layer these feature maps get downsampled, reducing in that way their dimensions while keeping all essential information. The refined features pass to the fully connected layers with a corresponding weight and then processed to classify the original input into the respective output category.

Convolutional layer has a significant role in CNN architecture and function. A convolutional layer applies several filters, also known as *kernels*, to input image and as a result feature maps are created through convolution operations. More specifically, filters slide over the input and calculate dot products between the filter weights and the inputs. With this process each filter detects features such as edges, textures or patterns. Usually, this output passes from an activation function such as the Rectified Linear Unit (ReLu) to add non-linearity into the model [4].

[4] Source: https://www.sciencedirect.com/science/article/pii/S0899707121002795

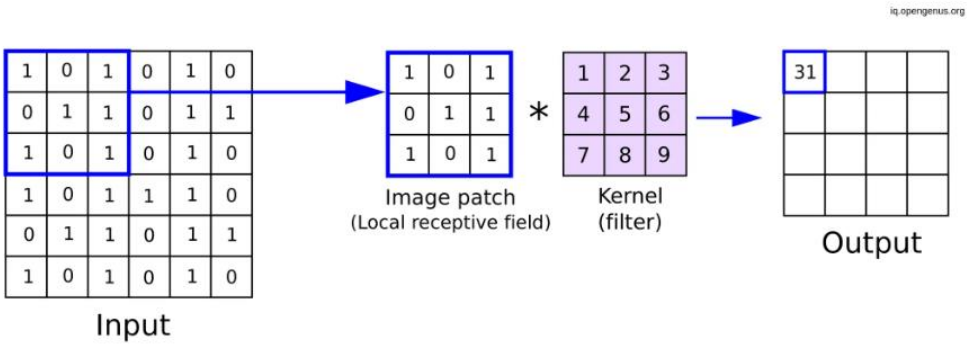


Figure 6. A visual representation of a convolutional layer [5]

Pooling layer reduces the dimensionality of the feature maps generated by the convolutional layer, while keeping all the important features. Through this downsampling process the model's computational complexity is reduced. Max-pooling layers are most used in CNN models where the maximum value form a specific area of activation map is selected. Typically, kernels used with these layers have small dimensionality (2x2). In this way the number of parameters that the network has to use is significantly reduced, enhancing model's efficiency [4].

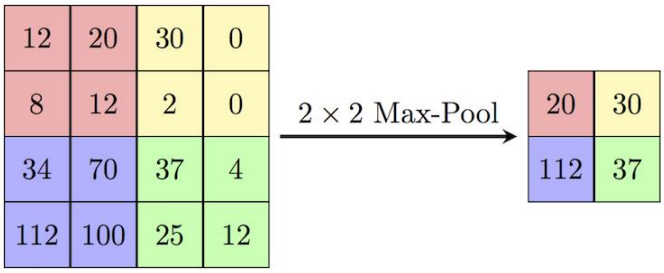


Figure 7. A visual representation of a max-pooling layer operation [6]

Each neuron in this layer is connected to each neuron in the previous layer, similar to the way that neurons are connected in traditional ANNs. The purpose of this layer is to process the extracted features from convolutional and pooling layers, in order to generate class score for classification tasks. For performance enhancement, a ReLU activation function may applied again in this layer [4].

---

[5] Source: https://www.superannotate.com/blog/guide-to-convolutional-neural-networks

[6] Source: https://paperswithcode.com/method/max-pooling

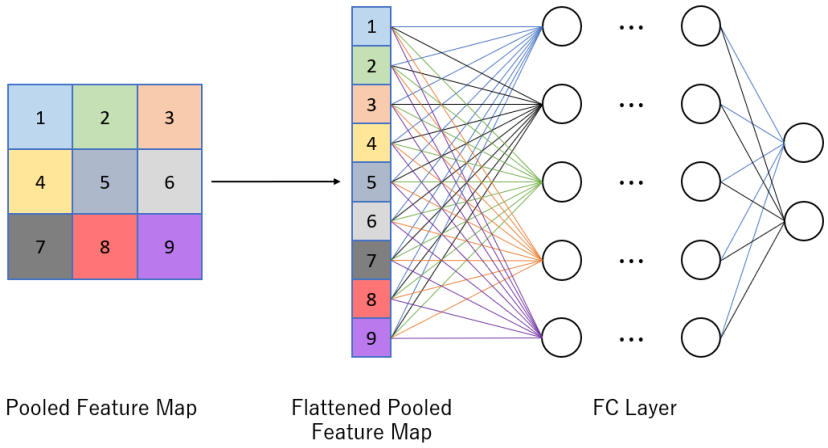


Figure 8. A pooled feature map being flattened and given as input to fully connected layer [7]

ResNet

Deep residual network also known as *ResNet* was introduced by He et. all in 2016 [5]. Authors created this architecture to address challenges in deep learning model training. Usually, training takes a lot of time and becomes more difficult as the number of layers increases. ResNet can handle these issues by skipping connections (making shortcuts). It skips connections by bypassing two or three layers that usually contain ReLU activation function and batch normalization. ResNet model has shown better performance in image classification tasks compared to other models, indicating excellent feature extraction capability. Authors proposed residual learning which is applied to multiple layers. The formula below defines the residual block:

$$y = F(x, W + x)$$

where x is the input, y is the output and F function represents the residual mapping learned by the block.

Inputs and outputs must have the same dimensions to work properly. In ResNet18 and ResNet34 each block has two layers while in ResNet50 and ResNet101 has three layers per block. Figure 9 illustrates and compares all different ResNet model variants [6].

[7] Source: https://medium.com/@alejandro.itoaramendia/convolutional-neural-networks-cnns-a-complete-guide-a803534a1930

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure 9. Architectural comparison of ResNet variants (18, 34, 50, 101, and 152 layers) [5]

EfficientNet

EfficientNet is a group of efficiently scaled convolutional neural networks designed to optimize both accuracy and efficiency through neural architecture search. Its architecture is based on the mobile inverted bottleneck convolution (MBConv) adapted from the MobileNetV2 [7]. EfficientNets have come out as state-of-the-art models in image classification tasks. The basic network in the EfficientNets family is the EfficientNet-B0 architecture as shown in Figure 10. MBConv module is the core structure of the EfficientNet-B0 and combined with Squeeze-and-Excitation (SE) blocks improve channel attention. Excluding the fully connected layer which feeds into a softmax classifier, each layer consists of a batch normalization and ReLU activation. Strided convolutions within the depthwise layers manage downsampling while a global average pooling is performed prior to the classification layer [7].

Stage i	Operator $\hat{\mathcal{F}}_i$	Resolution $\hat{H}_i \times \hat{W}_i$	#Channels $\hat{C}_i$	#Blocks $\hat{L}_i$
1	Conv3x3	224 × 224	32	1
2	MBConv1, k3x3	112 × 112	16	1
3	MBConv6, k3x3	112 × 112	24	2
4	MBConv6, k5x5	56 × 56	40	2
5	MBConv6, k3x3	28 × 28	80	3
6	MBConv6, k5x5	14 × 14	112	3
7	MBConv6, k5x5	14 × 14	192	4
8	MBConv6, k3x3	7 × 7	320	1
9	Conv1x1 & Pooling	7 × 7	1280	1
10	FC	1 × 1	1000	1

Figure 10. EfficientNet - B0 architecture

VGG

VGGNet was first introduced by Simonyan and Zisserman in 2014 [8]. The simplest CNN in the VGGNet family is the VGG-11. It is an 11-layer deep CNN that consists of 8 convolution layers and 3 fully connected layers and takes as input RGB images of 224x224 pixel size. These inputs are processed by convolutional layers of 3x3 kernel size and stride of 1. Max pooling layers are also employed between some of convolutional layers with a 2x2 kernel size and padding of 2. Moreover, the ReLU is used as an activation function in some of the fully connected layers [9][10][11].The figure below (Figure 11) illustrates the VGG-11 architecture.

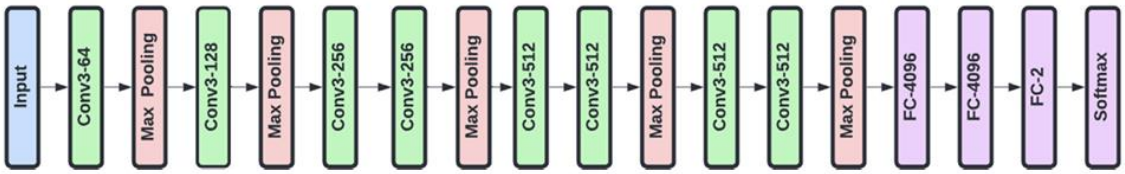


Figure 11. VGG-11 architecture

I3D

The Two-Stream Inflated 3D ConvNet (I3D) was first introduced by Carreira and Zisserman in 2017 [12], to improve action recognition in videos by capturing spatiotemporal features. Specifically, this architecture is based on successful 2D ConvNet designs, particularly the Inception-V1 model, since it enables the 3D inflation of its filters and pooling kernels (e.g., a 2D N×N filter becomes a 3D N×N×N filter) which, in turn, allows spatiotemporal feature extraction from videos. The core innovation is the inflation process, which allows the model to reuse the pre-trained weights from image classification tasks (e.g. ImageNet) and use them for video understanding. I3D model processes high resolution video sequences, typically 64 frames clips at 25 fps, which allows to capture high resolution temporal patterns in actions across all frames. In detail, the model is structured in two parallel streams, an RGB stream and a flow stream, separately trained, where their outputs are combined in order to enhance its performance. Moreover, the I3D's pre-training in Kinetics dataset significantly improves model's recognition accuracy, outperforming previous state of the art models in motion datasets such as UCF-101 and HMDB-51. Overall, the I3D model's inflating design enables complex spatiotemporal patterns learning, resulting in advanced recognition capabilities in video-based datasets [12].

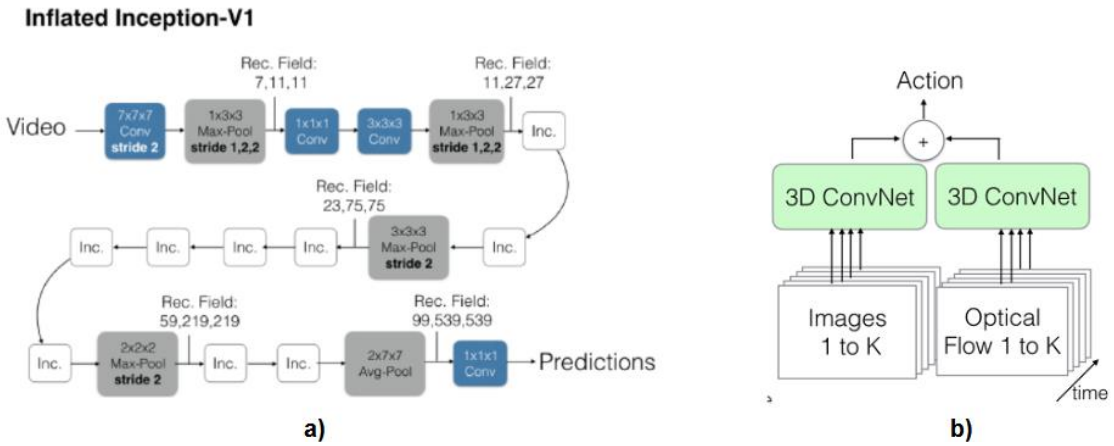


Figure 12. I3D Architectures: (a) Single-stream RGB Model (our implementation), (b) Full Two-stream Model (RGB + Optical Flow)

2.1.1.2 Transformers Neural Networks

Transformers are a type of neural network that uses a self-attention mechanism for feature extraction. Introduced by Vaswani et al. [13] in 2017 and first applied in natural language processing (NLP) tasks such as machine translation and text parsing. They have since attracted attention in computer vision tasks due to their capabilities, outperforming even CNNs in vision applications such as image classification. In their study Dosovitskiy et al. [14] using transformers achieved state-of-the-art results on multiple image recognition benchmarks datasets. Furthermore, transformers showed that they are capable of addressing multiple computer vision tasks including image segmentation, image processing, object detection and video processing [15]. A transformer consists of an encoder and a decoder networks including transformer blocks of the same architecture. The encoder creates encodings of inputs and the decoder takes all the encodings with their incorporated contextual information and outputs the sequence. Each transformer block consists of a multi-head attention layer, feed forward neural network a shortcut connection and layer normalization [15].

The self-attention mechanism is the core component of transformer architecture. The input vector converted into three separated vectors, a query (q), a key (k) and a value (v) with the same dimensionality, 512. These vectors are grouped together into three different matrices called Q, K and V.

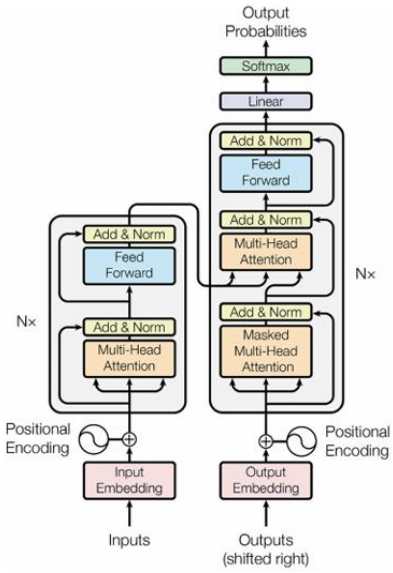


Figure 13. Transformer structure

Continuing, the attention function between input vectors is calculated. The below function shows the process unified.

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q} \cdot \mathbf{K}^\top}{\sqrt{d_k}}\right) \cdot \mathbf{V}.$$

The multi-head attention mechanism is used to enhance the performance of the self-attention layer. It allows the transformer to pay attention at the same time to information from different representation subspace. Instead of computing a single attention function, transformer can use multiple attention functions (*heads*) in parallel. In that way, each head can learn different relationships between input data [15].

Vision Transformer

Vision Transformer (ViT) is an architecture that uses the transformer model, which was first designed for NLP to image classification tasks. Images divided into fixed-size patches and flattened into vectors. Transformers blocks and a self-attention mechanism allow model to recognize patterns across different image patches in order to understand the entire image. ViTs when trained on large datasets, achieved state-of-the-art results outperforming CNNs trained on similarly large datasets. However, when trained on small datasets ViTs don’t generalize well,

compared to CNNs, due to lack of some inductive biases [15]. The figure below (Figure 14) demonstrates the ViT framework.

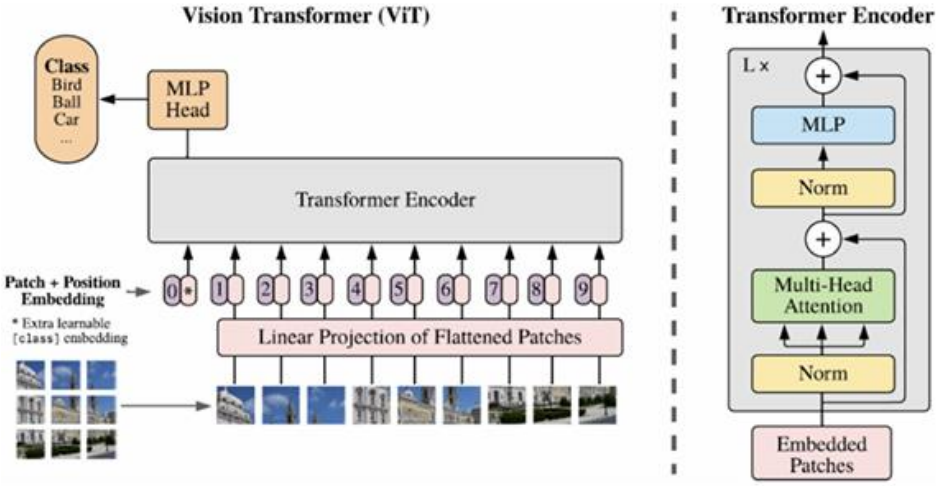


Figure 14. ViT framework

2.1.2 AI challenges in medical imaging

The wide adoption of AI in medical imaging has several challenges [16]. Firstly, there is data heterogeneity and complexity issue which arises from different image formats, resolutions and protocols. For instance, most medical images are stored in Digital Imaging and Communications in Medicine (DICOM) format which offers a variety of resolutions and bit depths. On the other hand, datasets such as *ImageNet* consist of low-resolution JPEG images and these images are totally different from medical images. Also, medical images such as MRI and CT generate multiple image types and views. This issue makes the development of AI models more complex than developing an AI model for non-medical images. Secondly, data curation for medical images is expensive and susceptible to observer bias, since trained radiologists are needed to label medical images. In comparison non-medical images labelling is much easier since most of the objects illustrated are easily recognized. Another critical challenge is patient privacy. DICOM images contain protected health information (PHI) in metadata and medical images can contain PHI even in the pixel data itself. Moreover, patients' accessories including necklaces and wristbands may contain unique identifying information such as patient's name, enough to allow patients to be recognized through the images. This makes the de-information process expensive and labor-intensive. Performance evaluation metrics selection is critical because metrics such as

accuracy can be misleading in imbalanced data. This means that accurate and clinically safe AI models must be developed, which makes constructing the algorithms even more challenging. In addition, AI models lack transparency and explainability in their decision-making processes, which further complicates their validation. Many models work as 'black boxes' with limited understanding of their reasoning. One more challenge of AI in medical imaging is model generalizability and transferability. Usually, models are trained on datasets with specific characteristics, however, patient populations vary between clinical institutions, leading to poor generalizability. Differences in imaging protocols and demographic data make transferability very difficult. Moreover, despite the high accuracy of AI models and their promising results in medical imaging, the aforementioned difficulties slow down their integration in clinical workflow [16].

2.1.3 Pre-processing Techniques

In this section we describe the pre-processing techniques employed prior model training in this study. To improve deep learning models' performance, pre-processing techniques were applied as an essential step in both tasks. As all processes, these techniques have additional computational cost, however, their application is critical for achieving reliable and improved classification. Each task requires different approaches in terms of data preparation and model selection while their results are evaluated with the same metrics. In this work, the pre-processing phase includes several stages such as data augmentation, image cropping, data splitting, image resampling, normalization, standardization and harmonization techniques that are further described below.

Data Augmentation

Data augmentation is a set of techniques that artificially increase and enhance training datasets by applying transformations, such as geometric adjustments, color changes and the use of generative models like GANs. It mainly focuses on addressing the challenge of limited data availability, which is especially critical in domains such as medical imaging, where data collection is expensive, time-consuming and often limited by privacy concerns. Data augmentation reduces overfitting and improves deep learning model's generalization by increasing the diversity and size of the training dataset. More specifically, in medical imaging, the collection of accurately labelled datasets is difficult due to the high cost of expert annotation and patient privacy restrictions, making augmentation techniques such as GANs crucial due to their significant effectiveness. Such methods have been shown to generate realistic and diverse images that enhance classification sensitivity and specificity, which improves the overall accuracy and reliability of diagnostic models that process medical images [17]. The below figure (Figure 15) illustrates some geometric augmentations that applied to a COVID-19 X-ray image.

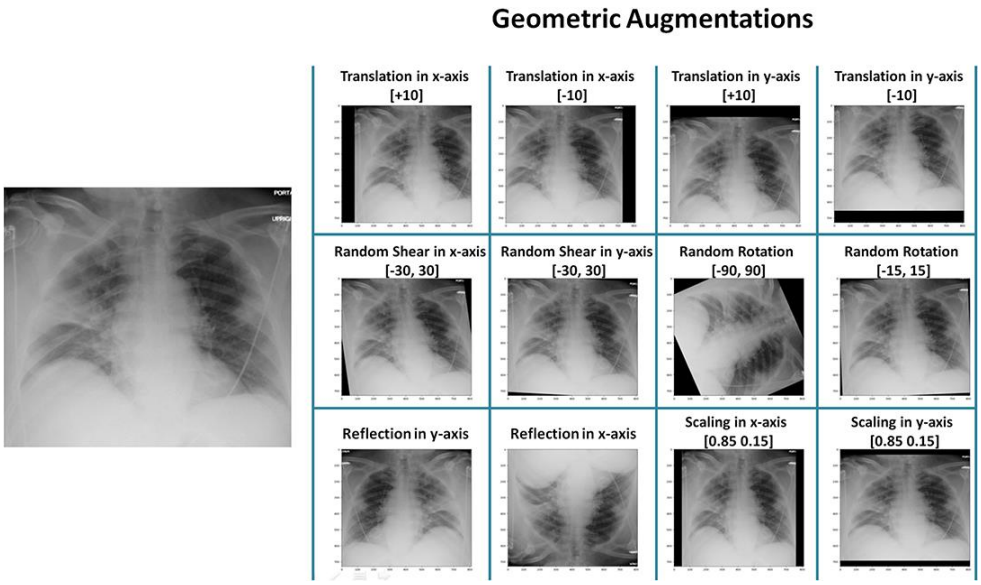


Figure 15. Examples of geometric data augmentation on chest X-ray image [8]

Image Cropping

Image cropping can be used as a pre-processing step and belongs to data augmentation techniques. It is a useful method especially when there is a set of images with mixed dimensions (height and weight), that effectively reduce their dimensions by removing part of images. It's very valuable in input standardization for neural networks. However, this technique is facing some challenges, especially if the removed parts are essential for correct classification, leading to poor model performance. The below figure (Figure 16) shows how image cropping reduces the spatial dimension of chest X-ray image, by removing peripheral regions, resulting in a more appropriate input size for neural network processing [17].

[8] Source: https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2021.629134/full

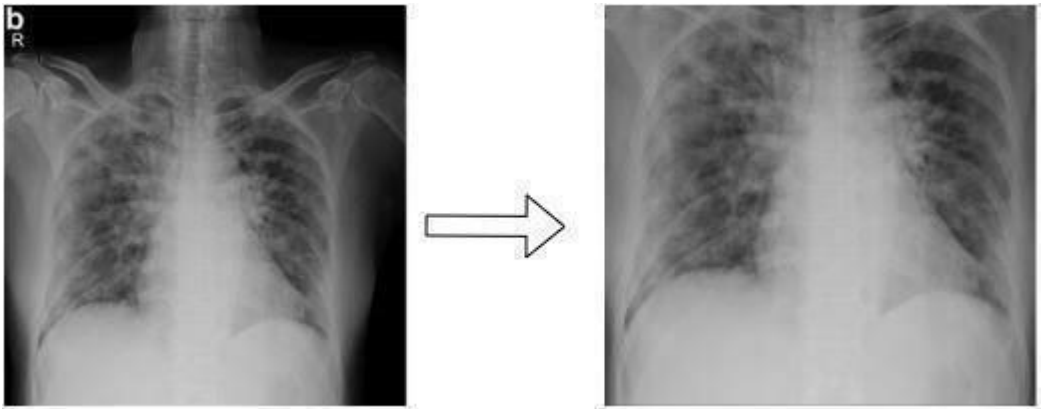


Figure 16. Example of image cropping on chest X-ray [9]

Data Splitting

In the field of machine learning, specifically in supervised tasks, data splitting refers to the process of partitioning the available dataset into three distinct subsets, training, validation and test, each having a specific role in model development. The training set is a set of example data that is used to fit model parameters such as weights in neural networks, in order to learn from input examples. The validation set is a set of example data that is used by the fitted model to evaluate its performance in unseen observations. Also, it's used for model's hyperparameters tuning and overfitting prevention. Lastly, the test set is an unseen set of data, that is used for final performance evaluation; to figure out how well the model performs to new, unseen data. Often, it is noticed in literature, that the term "validation set" is used instead of the term "test set" because the original dataset is divided into two subsets[10].

Resizing & Resampling

Resizing is the process of changing the spatial resolution of an image, by scaling it up or down (e.g. making it smaller or larger). In the field of computer vision image resizing is a common pre-processing method that prepares images for efficient processing by deep learning models which often required fixed input size (e.g. 224x224). **Resampling** refers to image resizing processes that use interpolation techniques such as bilinear, bicubic or Lanczos. Specifically, these

[9] Source: https://www.researchgate.net/figure/An-example-of-Cropping_fig2_343264351

[10] Source: https://en.wikipedia.org/wiki/Training%2C_validation%2C_and_test_data_sets

methods estimate new pixel values based on the surrounding pixels. In that way, image resolution is changing while the visual parts of the image remain maintained [18].

Normalization

Min-max normalization, also known as feature scaling, is a widely used normalization technique that is used to rescale the intensity values of an image (e.g. MRI). The scope of this technique is to rescale the intensity values into a specific range which is usually between 0 and 1. In order to perform this method, the minimum and maximum intensity values of the image must be identified. Each intensity is then rescaled by subtracting the minimum and dividing by the overall intensity range (i.e. the difference between the maximum and minimum). As a result, this process rescales all the intensities into a 0-1 range [19].

Standardization

Standardization, also known as z-score, is a widely used technique in the field of machine learning. It rescales the data to have a mean of 0 and variance of 1. Specifically, each feature value is adjusted by subtracting the feature's mean and then dividing by its standard deviation [20][21].

2.1.4 Harmonization Techniques

Nowadays, it is noticed that the amount of data in the field of medical imaging is constantly increasing. Data is acquired from different equipment, acquisition protocols and clinical centers. The integration of multi-center and multi-device data can enhance studies generalizability and statistical power of results. However, a critical challenge arises, factors such as differences in pixel spacing, imaging parameters, image intensity and hardware equipment variations can introduce inconsistency in images and reduce comparability or lead to domain shift problems.

Image harmonization is a technique that aims to overcome inter-site variation and enable reliable and meaningful comparisons and analysis across different imaging settings [22][23]. Harmonization can be performed through image normalization methods also known as pre-processing, standardization, normalization or denoising. Employing harmonization techniques in input data, allowing AI models to successfully focus on key image features instead of being distracted by differences arising from different scanner settings/protocols or any other factors [22].

Figure 17 illustrates a pipeline that is usually used in multi-center studies. Images are acquired from different clinical centers and scanners pooled together and get harmonized. After that, they are used for AI model training.

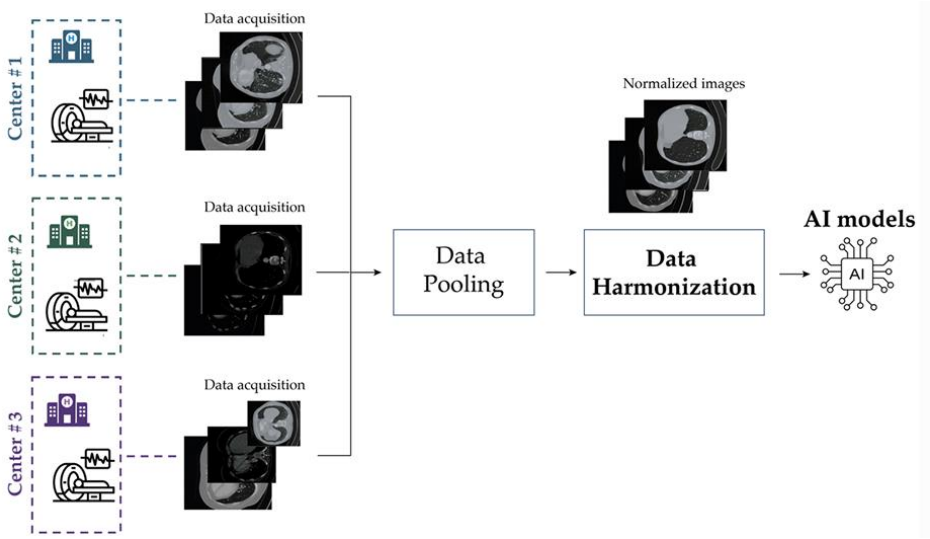


Figure 17. Common pipeline used in multi-center studies

For the purposes of this study, regarding the PCa management task, we employed four normalization methods widely used for normalization or harmonization: 1) Z-score normalization, 2) Histogram Normalization, 3) Mean Normalization, 4) Combination of Z-score & N4 Bias Field Correction.

Z-score Normalization

This technique is also known as zero mean unit and it can be applied fast and easily. It rescales and shifts the intensities using the formula:

$$I_{norm}(x) = \frac{I(x) - \mu}{\sigma}$$

where I is the intensity distribution, μ is the mean intensity and σ is the standard deviation of the whole image [24].

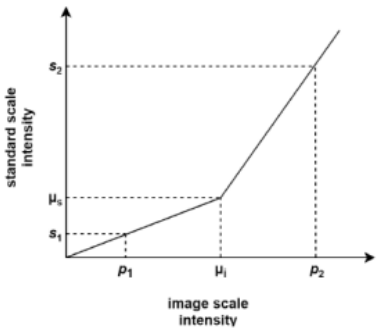
Histogram Normalization

It was first proposed by Nyul and Udupa in 1999 [25] and is one of the most successful normalization methods still today. Specifically, this method aligns the intensity histogram of the images with a standard histogram, which is established by a training process. This is done in a piece-wise linear manner, that ensures that the selected landmarks points such as mean or median intensity are mapped to standard scale on the same values. This technique first identifies the image's foreground in order to perform pixel isolation from the background. This process is achieved by thresholding the image based on the average intensity of pixels across all slices of a scan. On the foreground the training procedure of normalization is performed, and its mode is selected as the median landmark. Moreover, two other landmarks can be defined as the 0th and 98th percentiles of the entire histogram.

Continuing, during the transformation process, these three landmarks are extracted from the intensity histogram and successfully are mapped into the standard scale landmarks using a piecewise linear transformation. These landmarks are determined during the training phase that includes multiple scans of a body region and imaging protocol.

S_1 and S_2 standard scale landmarks are specifically chosen to ensure one-to-one mapping in order to avoid histogram compression. Finally, the foreground mode is now mapped onto this new standard scale while the average remapped mode of the training scans is used as the standard scale median landmark. The below figure (Figure 18) illustrates the piecewise linear mapping. Landmarks intensities μ_i from different images on the original intensity scale are mapped to the same value μ_s on the standard scaler. Different linear functions map the intensities below and above landmark value[11].

[11] Source: https://www.procancer-i.eu/wp-content/uploads/2022/12/ProCAncer_D5.2_Pre-processing-tools_FORTH-vf.pdf



$$x' = \begin{cases} \mu_s + \frac{s_1 - \mu_s}{p_{1i} - \mu_i}(x - \mu_i), & \text{if } x \leq \mu_i \\ \mu_s + \frac{s_2 - \mu_s}{p_{2i} - \mu_i}(x - \mu_i), & \text{if } \mu_i < x \end{cases}$$

Figure 18. Illustration of the piece-wise linear mapping [11]

Mean Normalization

Arguably mean normalization is one of the simplest normalization techniques. It rescales the image intensity values by subtracting the mean value of the whole MRI scan[11].

$$x' = x - \mu$$

Combination of N4 Bias Field Correction and Z-score Normalization

N4 Bias Field Correction method is an improved version of the successful N3 Bias Field Correction. It is being applied to images to remove bias field artifacts and it is highly recommended to be used as a pre-processing step in multi-center studies [24]. In this study, we selected to employ the combination of both techniques because it is noticed that N4 Bias Field Correction followed by Z-score normalization can significantly increase model performance compared to N4 method alone [26].

2.1.5 Evaluation Metrics

For model's performance evaluation we mainly employed balanced accuracy, AUC score and specificity, as these metrics are especially useful in the case of imbalanced datasets. In addition, to provide a comprehensive overview we present the confusion matrix, a 2x2 (in binary classification tasks) table that visualizes the actual and the predicted instances per class by showing four values TP, TN, FP, FN [27][28] as shown in Table 1. While our main analysis focuses on the aforementioned metrics, we briefly describe additional evaluation metrics including accuracy, sensitivity, precision and F1-score to provide a better understanding of how model performance can be assessed.

Table 1. Confusion Matrix

		Predicted Class		**Total Instances**
		0	**1**	
Actual Class	0	True Positive (TP)	False Negative (FN)	P
	1	False Positive (FP)	True Negative (TN)	N

- True Positives (TP): The number of instances that have classified correctly as '0'
- True Negatives (TN): The number of instances that have classified correctly as '1'
- False Positives (FP): The number of positive instances that are actual '1' but have classified as '0'
- False Negatives (FN): The number of negative instances that are actual '0' but have classified as '1'

Using the above values of the confusion matrix we can easily calculate the metrics accuracy, sensitivity, specificity, precision and f1-score.

Accuracy describes the instances that model have correctly classified:

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

Sensitivity (also Recall or true-positive rate) calculates the proportion of positive instances that are correctly predicted as positives:

$$\textbf{Sensitivity} = \frac{TP}{TP+FN}$$

Specificity (also true-negative rate) calculates the proportion of negatives instances that are correctly predicted as negatives:

$$\textbf{Specificity} = \frac{TN}{TN+FP}$$

Precision calculates the proportion of actual positive instances from all cases that the model predicted as positives:

$$\textbf{Precision} = \frac{TP}{TP+FP}$$

F1-score (also F-score or F measure) is the harmonic mean of precision and sensitivity:

$$\textbf{F1-score} = \frac{2TP}{2TP+FP+FN}$$

Balanced Accuracy measures the average of sensitivity (recall) and specificity. It's useful for imbalanced data, where the traditional accuracy may be misleading [29]:

$$\textbf{Balanced accuracy} = \frac{sensitivity+specificity}{2}$$

The Receiver Operating Characteristic (ROC) curve is a graphical plot that displays the performance of a model by showing the trade-off between the true-positive rate and false-positive rate. The closer the ROC curve to the upper left corner, the more accurate the model is. The area under the ROC curve (AUCROC) shows overall how good the model is despite the threshold. However, a specific threshold is needed to define how a model operates in real-world scenario. AUCROC of 1 (best-case scenario) shows a perfect model performance while an AUCROC of 0.5 (worst-case scenario) shows that model performance is equivalent to random guessing [28].

2.2 Prostate Cancer Management

2.2.1 Introduction

Prostate cancer (PCa) is one of the most frequent cancers (after lung cancer) which affects middle-aged men between 45-60 years of age worldwide, with elderly people over 65 years be in higher risk. Based on estimates, in 2018, over 1 million cases of PCa cases reported in developed countries and over 2 million new cases are estimated until 2040 [30]. Worth noting that, African-American men incidence rates, are higher and they are more likely to develop more aggressive type of PCa compared to white men. However, PCa cases over the world are highly variable. More specifically, Europe, North America and Australia have higher incidence rates compared to Asia men population. Currently, it is not clear why these differences occur but dietary factors among countries worldwide are mainly associated. Moreover, the family history of PCa is the primary risk factor for the disease. Men with close relatives who have been diagnosed, face a 50% higher risk of developing PCa compared to those without any family history of it. Physical activity has also an important role in cancer development and progression [30][31].

At early stages, PCa may be asymptomatic without causing any pain to patient. Nonetheless, the most common symptoms are nocturia, difficulty with urination and increased frequency, and other symptoms that arise from prostatic hypertrophy. Patients in more advanced stages may experience urinary retention and back pain since the axis skeleton is the primary site for bone metastasis [30]. Currently, there is not a specific method/test for PCa diagnosis. Prostate-specific antigen (PSA) is a widely used test for PCa detection. Only 1 in 4 men with elevated PSA test will be diagnosed with PCa, at the same time patients with PSA< 1ng/ml still have 10% chance to develop the disease [32][33]. Usually, patients with elevated PSA are guided by their doctor to perform prostate biopsy [32]. In this procedure, a thin needle is used for tissue sample collection from the prostate gland for further examination under the microscope.

Magnetic Resonance Imaging (MRI) or Transrectal Ultrasound (TRUS) is used to locate the prostate gland. Biopsy analysis is among the most reliable methods for diagnosing PCa providing clinicians with a comprehensive view of the cancer's progression. In recent years, Artificial Intelligence (AI) algorithms have shown promising results, providing new classifications for PCa. Furthermore, advanced imaging techniques such as multiparametric Magnetic Resonance

Imaging (mpMRI) could lead to personalized diagnosis with recent guidelines to suggest an MRI scan before any prostate biopsy, minimizing in this way further complications [31].

Treatment options for PCa depend on PSA level, tumor nature, grade and stage. Available options include chemotherapy, radiation therapy, hormonal therapy, surgery, cryotherapy and monitoring. For instance, in disease monitoring, there is no treatment given to patients unless they become symptomatic while radical prostatectomy and hormonal therapy are used depending on tumor's stage. However, all treatment options can lead to severe side effects [31][34].

Even though there are numerous ways to diagnose PCa, there are some limitations. Biopsy is usually referred as the 'gold standard' for determining tumor aggressiveness. However, this procedure is invasive and uncomfortable for the patient [35]. As mentioned above, MRI is suggested before any prostate biopsy and has been shown that it is able to detect a greater number of aggressive cancers, reducing at the same time the number of biopsies and overdiagnosis. However, interpreting MRI requires significant expertise and there is a shortage of experienced readers. This limitation can lead to loss of up to 30% of tumors [36]. PCa due to clinical heterogeneity and the differences in prognosis, have prompted efforts to develop risk stratification methods for PCa management [37]. Computer-Aided Diagnosis (CAD) systems have been developed using AI in order to help radiologists by automating the detection of PCa on MRI scans. Indeed, clinicians support that AI systems could really improve their work but most of them have never used any type of AI systems, while only a minority considered themselves to have excellent knowledge of AI. Barriers to adoption in clinical practice such as lack of trust, knowledge of use, limitations and capabilities of these systems remain a big challenge [35].

2.2.2 Current advances in AI applications for PCa management

PCa clinical heterogeneity and high number of overdiagnosis/overtreatment has led the AI research community to develop risk stratification models for PCa management. Due to data privacy policies, training of such AI models using massive amounts of data (e.g. MR images) can be limited. However, data anonymization protocols in addition to open data sharing policies stimulated the creation of large clinical datasets. Datasets composed of different sources such as different scanners manufacturers and settings are now available [38]. Due to the heterogeneity that

these data present, pre-processing steps are necessary before model training [38]. Recent research underscores the impact of image pre-processing on deep learning analysis of prostate MRI.

Wu [39] normalized 500 MRI scans of benign, hyperplasia, prostate cancer and healthy males before CNN model training, achieving a mIoU of 79.4%. PM et al. [40] converted MRI images from DICOM to TIFF, followed by rescaling, intensity normalization and data augmentation, achieving 93.97% and 99% accuracy and AUC for PCa classification. Zaridis et al. [41] proposed a U-Net pipeline using Prostate-X dataset for peripheral-zone segmentation, while in a later study [42] introduced RACLAHE image enhancement method, outperforming four established contrast enhancement techniques using different CNN models. Saha et al. [43] employed intensity normalization, resampling and augmentation on bi-MRI images (T2W, ADC, DWI) to train a dual 3D-CNN CAD system, reaching an AUROC of 82.20%. In a similar study, Ali & Mohammed [44] applied a simple pre-processing pipeline using add/remove noise algorithms, image resizing and DICOM to JPG format conversion, with their 2D CNN-based approach achieving AUC of 98% without pre-processing and AUC of 99% with pre-processing methods applied, in PCa classification. Liu et al. [45] addressed data heterogeneity in Prostate-X dataset, achieving 84% AUC score with a custom 3D CNN model. Sherafatmandjoo et al. [46] combined mpMRI data with clinical and pathological features and they achieved average accuracy 85% without and 95% with clinical pathological data respectively. Yoo et al. [47] combined ResNet models and a Random Forest using 5,832 DWI images, with the CNN pipeline achieving AUC of 87% and Random Forest 84%. Aldoj et al. [48] used mp-MRI images from Prostate-X dataset investigated the feasibility of using deep learning techniques for semi-automatic prostate cancer lesions classification. They achieved an average AUC of 89.7% using a 3D CNN. Vente et al. [49] used ProstateX-2 focusing on detection and grading PCa in bi-parametric MRI scans. Their approach achieved a kappa score of 0.446 and Dice 0.37 for PCa aggressiveness grading. Hassan et al. [50] compared several pre-trained models using ultrasound and MRI data, with Xception model achieving the best performance with 97% accuracy on ultrasound images and 80% on MRI images. Lastly, Pachetti et al. [51] after applying data normalization and augmentation on T2W scans from ProstateX-2, trained a 3D Vision Transformer achieved a mean AUROC 92.7% outperforming a 3D CNN model (mean AUC of 58.5%).

Open issues

In the context of prostate cancer management using AI, different methodologies, datasets and tools have been used for PCa prediction. Deep learning has shown high accuracy in PCa management, particularly in classification and segmentation tasks. MRI images including T2W, DWI and ADC sequences proved that are extremely valuable in supporting AI models to analyze prostate abnormalities. Several studies employed advanced AI architectures based on 2D/3D CNNs and ensemble models using at the same time techniques such as data augmentation, normalization, noise add/removal and image resampling as pre-processing steps in order to enhance model performance. Moreover, the integration of MRI images with pathological and clinical data or ultrasound images showed improved model performance compared models trained only on MRI images.

However, several challenges still exist. It was observed in literature that most of the pre-processing techniques employed are used in an ad-hoc fashion and there is no clear evidence regarding their impact on AI model performance. Limitations in datasets such as small size and population homogeneity, further limit models' generalizability. Moreover, broader applicability of these models is restricted due to the lack of external validation and vendor-specific imaging reliance. Image data heterogeneity, arising from different MRI vendors, machines, acquisition protocols remain a critical challenge. So pre-processing pipelines and harmonization methods are necessary. Image pre-processing and harmonization before model training have shown that it can assist in solving such challenges and significantly enhance AI models' performance.

2.2.3 Notable Datasets

In the literature on PCa management [41][42][45][48][49][51], two publicly available datasets—PROSTATEx and PROSTATEx-2—are among the most frequently used. These datasets were released through the *SPIE-AAPM-NCI Grand Challenge* and they have been established as benchmarks datasets for prostate MRI methods diagnosis and computer-aided detection evaluation. Both datasets are accessible through The Cancer Imaging Archive (TCIA)[12].

[12] Source: https://www.cancerimagingarchive.net/collection/prostatex/

PROSTATEx dataset was first released in 2017 as part of *SPIE–AAPM–NCI Prostate MR Classification Challenge*, and still today it is one of the most used public resources for development and benchmarking prostate MRI analysis methods. It includes multiparametric MRI (T2W, ADC, DWI and dynamic contrast enhanced (DCE) MRI sequences) of 538 prostate lesions acquired by 344 patients using Siemens 3T scanners (MEGNETOM and Skyra). All images were annotated by expert radiologists, labeled as *not clinically significant* and *clinically significant* and the cancer ground truth was established through targeted biopsies [52].

As a follow up to PROSTATEx, PROSTATEx-2 dataset was released in 2019 as part of *SPIE-AAPM-NCI Prostate MR Gleason Grade Group Challenge*. However, it was designed for a more challenging task, Gleason Grade Group (GGG) prediction of each lesion. MRI sequences remained the same as the PROSTATEx dataset, only the labelling was changed. Specifically, in this dataset each lesion was assigned to a GGG score from 1 to 5, with a total of 182 lesions across acquired from 162 patient cases [52].

2.3 Therapy-Induced Cardiac Toxicity

2.3.1 Introduction

According to Centers for Disease Control and Prevention (CDC) breast cancer is the second most common cancer among women in the United States, with over 279,000 new cases in 2022 and 42,000 deaths in 2023[13][14]. Nevertheless, it has been noticed that among breast cancer survivors > 65 year of age, cardiovascular (CV) disease is the largest cause of death. Despite cardiovascular risk factors such as age, diet and lifestyle, several cancer treatments including chemotherapeutics, radiation therapy and molecular targeted therapies can cause cardiac side effects. According to the National Cancer Institute (NCI), cardiotoxicity is defined as "toxicity that affects the heart" [53][54]. Breast cancer therapy-induced cardiotoxicity can cause heart failure, asymptomatic left ventricular (LV) disfunction, arterial hypertension, pericarditis, thromboembolic events, myocardial ischemia and arrythmias. Chemotherapy plays an important role in breast cancer treatment. Anthracyclines is a widely used drug to treat breast cancer, but it has been shown that it can cause acute and chronic cardiovascular effects such as myocardial injury, months or years after the drug exposure [53][55]. Additionally, trastuzumab is an approved and commonly used chemotherapy in breast cancer treatment, where the protein HER2 is overexpressed. However, this therapy is associated with cardiac dysfunction such as cardiomyopathy or congestive heart failure (CHF) [56].

Cardiac function should be assessed before, during, and after anticancer treatment. Techniques including strain echocardiography, 3D echocardiography, electrocardiogram (ECG), cardiac magnetic resonance imaging (CMRI), combined with the use of biomarkers allows early detection of cardiac damage [57][58]. Echocardiography is a noninvasive and low cost widely used method for cardiac evaluation of cancer patients. It provides a simple way to assess Left Ventricular (LV) systolic function, particularly when combined with the Left Ventricular Ejection Fraction (LVEF) or strain. For the diagnosis of early LV dysfunction in an asymptomatic patient, LV strain may be recommended over LVEF. For patients undergoing trastuzumab therapy echocardiography follow-up is recommended at baseline and every three months to monitor LV

[13] Source: https://www.cdc.gov/united-states-cancer-statistics/publications/breast-cancer-stat-bite.html

[14] Source: https://www.cdc.gov/breast-cancer/statistics/index.html

dysfunction. LVEF is the most essential metric for diagnosing and evaluating HF with the current echocardiographic guidelines for cardiotoxicity are a decrease in LVEF by 10% point or a reduced LVEF < 53% [58].

Challenges

Early detection and management of cardiotoxicity are vital, preventing any irreversible cardiotoxicity [59][60]. Advanced imaging methods including CMRI, 2D echocardiography, ECG and AI models for identifying cardiotoxicity show limitations. 2D echocardiography and ECG are commonly used for cardiotoxicity prediction, however, they lack detection sensitivity in the beginning of myocardial damage [59]. Specifically, echocardiography diagnostic accuracy is limited and relies on operator's expertise and skills, which may lead to bad reproducibility compared to other imaging techniques. Also, image quality may be poor in some patients [58][60]. Despite the advanced AI technologies, clinicians are still cautious about their use and adaptation in clinical practice. Main limitations include concerns about the clarity and validation of risk scores generated by AI, making their application and interpretation complicated. Ethical dilemmas arise regarding the responsibility for monitoring AI-generated outcomes and ensuring timely interventions, raising questions about the consequences of depending on AI for decision-making. Furthermore, the adoption of AI in clinical workflows remains limited with some clinicians considering these technologies unnecessary, highlighting the importance of human judgment in the final decision-making [61]. However, despite the aforementioned limitations, as illustrated in Figure 19, AI is used in various cardiac applications, including assessment of cardiac function, diagnosis and prognosis of cardiomyopathy, prognosis of coronary artery disease and plaque analysis. Detection, segmentation and classification are the most common tasks of AI algorithms in cardiac imaging [62]. Taking advantage of the increasing computing power and the availability of cardiac image datasets, the research community can now develop more robust cardiac analysis models, discover new clinical insights on cardiac health and disease, developing at the same time novel AI tools assisting in clinical cardiology [63].



Figure 19. Four main categories of AI tasks in clinical cardiac radiology [62]

2.3.2 Cardiotoxicity risk assessment

Liu et al. [64] presented an end-to-end deep learning framework that distinguishes between four similar cardiovascular diseases and healthy individuals using apical-4-chamber (A4C) echocardiographic videos acquired from 1276 patients. Their approach achieved comparable performance to expert cardiologists, with AUC scores over 98% in all categories. Chang et al. [65] through their study aimed to differentiate patients at high risk of developing cardiotoxicity specifically cancer therapy-related cardiac dysfunction (CTRCD) and symptomatic heart failure with reduced ejection fraction (HFrEF). Authors applied several machine learning algorithms and used clinical data from breast cancer patients undergoing anthracycline therapy, achieving an AUC of 66% and 79% for CTRCD and HFrEF respectively. In a similar fashion, Choi et al. [66] presented a hybrid approach that combined expert knowledge with machine learning algorithms to assist clinicians in heart failure diagnosis. Their approach reached 100%, 100%, 99.5% and 91.7% accuracy for HFrEF, heart failure with mildly reduced ejection fraction (HFmrEF), heart failure with preserved ejection fraction (HFpEF) and no heart failure respectively. Madani et al. [67] trained a CNN with echocardiograms based on multi-vendor videos and static images data,

collected in a 17-year period. The model achieved an average accuracy of 97.8% and 91.7% on videos and static images respectively. Magesh and P.S RajaKumar [68] focused on congenital heart disease (CHD) in the fetal heart. They developed a CNN for normal and abnormal fetal heart classification based on ultrasound images obtained from 363 pregnant women. Their approach achieved 98.4% for normal and 97.2% for CHD. Zhang et al. [69] trained multiple CNNs in different tasks, classification, segmentation and disease detection with over than 10,000 echocardiograms from 23 different echocardiographic views. They reached high classification accuracy with C-statistics of 93% for hypertrophic cardiomyopathy, 87% for amyloidosis, and 85% for pulmonary hypertension. Comparative work by Howard et al. [70] used over 8,000 echocardiographic videos in 14 different echocardiographic views to explore the effectiveness of state-of-the-art CNN architectures in ultrasound video classification. The CNN architectures included classic, classic encapsulated within a time-distributed layer, 3D and two stream CNNs. Their results showed that two-stream and time-distributed CNNs achieved error rates as low as 3.9% with Cohen's Kappa of 0.957. In contrast, Talebi et al. [71] proposed a machine learning approach based on radiomics for radiotherapy-induced cardiotoxicity prediction in breast cancer patients (with vs without cardiac toxicity). Authors employed a variety of clinical and demographic data as well as ultrasound imaging radiomics features and performed four different experiments. The category that included dosimetric, demographics, clinical and imaging data achieved the best performance with AUC of 97%. Tsiouris et al. [72] presented a prediction and risk assessment model of early cardiac dysfunction in older breast cancer patients over 55 years, based on diverse baseline factors using a multi-center center dataset consisting of 1,560 breast cancer patients treated with chemotherapy. LVEF decline used for risk evaluation, categorized patients into two categories, LVEF decline and No LVEF decline. Results showed that the accuracy of LVEF decline prediction was improving during follow-up months horizon, starting from 66.67% accuracy (month 3 after baseline) and increased up to 73.55% (month 12). Lastly, Ouyang et al. [73] presented an end-to-end DL approach called *EchoNet-Dynamic* for the left ventricle segmentation and ejection fraction estimation, from echocardiogram videos. Authors employed 10,030 A4C view echocardiography videos collected by expert sonographers in a 2-year period. Their approach predicted the ejection fraction with a mean absolute error of 6.0% and classified heart failure with reduced ejection fraction with AUC of 96% on external dataset.

Open issues

Application of AI in cardiotoxicity risk assessment and related cardiac imaging task have demonstrated promising results, with models showing accurate detection, classification, often reaching expert-level performance using various data types.

However, several challenges persist. Data heterogeneity and small datasets limit models' generalizability. Lack of external validation based on different populations or healthcare systems in current studies remains an open challenge. Furthermore, clinicians are still hesitant to use and trust AI models in clinical practice due to poor explainability, especially in deep learning models, usually referred as "black boxes". It has been noticed in literature that there is lack of AI models for predicting presence of cancer therapy-induced cardiotoxicity in breast cancer patients undergoing cancer treatment using echocardiography video sequences data. To the best of our knowledge, there are no other attempts to predict cardiotoxicity from image data. Through this thesis we aim to shed light on the utility of 3D AI models to assist in breast cancer induced-cardiotoxicity risk prediction using 3D ultrasound video data.

2.3.3 Notable datasets

For the cardiotoxicity risk prediction task, in [73] study, EchoNet-Dynamic[15] is the closest publicly available dataset to our case. Although it does not consist of cardiotoxicity data acquired from breast cancer patients undergoing cancer treatment, it focuses on beat-to-beat LV function, making it suitable for benchmarking models that analyze cardiac motions from heart video sequences. It is a publicly available dataset of 10,030 A4C echocardiography videos obtained by individuals in a two-year period (2016-2018) as part of routine clinical care with each video to correspond to a unique patient and unique visit. According to authors, a standard full resting echocardiogram study consists of videos and static images visualizing the heart from different positions and angles, obtained with a variety of acquisition techniques. Specifically, expert sonographers obtained the videos using iE33, Sonos, Acuson SC2000, Epiq 5G or Epiq 7C ultrasound machines. Each video was cropped and masked in order to remove irrelevant information outside the scanning sector and standardized into 112x112 pixel resolution. Each

[15] Source: https://echonet.github.io/dynamic/index.html#dataset

video was linked with clinical measurements such as the left ventricular end systolic volume (ESV), left ventricular end diastolic volume (EDV) and the ejection fraction (EF) calculated by (EDV - ESV) / EDV formula. Moreover, expert annotations include two frame-level endocardial tracings of the left ventricle, one at end-systole and one at end-diastole encoded as pair coordinates. Each tracing is used for ventricular volume estimation by integrating the ventricular area over the length of the major axis of the ventricle [73].

3. Harmonization techniques and their impact on clinically significant PCa classification

3.1 Introduction

In this section we provide a detailed description of the implementation process for the first task of this thesis, PCa management using AI, as illustrated in Figure 20. A part of ProstateNET dataset that includes T2W, ADC and DWI MRI sequences in 32x224x224 volume size followed by their clinical data was used. For the pre-processing step, a variety of harmonization techniques were employed to handle multi-centric and multi-vendor data. From the original dataset, two different datasets were generated (24x192x192 and 24x128x128 volume size) to experiment models' performance using lower resolution images and reduced depth, with the harmonization techniques applied to all datasets. Moreover, several data augmentation transforms applied to increase the diversity of the training data and improve models' generalization. Subsequently, four 3D DL models were trained on the above data, EfficientNet, ResNet18, VGG11 and a Vision Transformer. Additionally, from clinical data, patient information such as age, PCa aggressiveness (non-clinically and clinically significant) and MRI scanner settings including magnetic field (1.5T or 3T), manufacture and model were extracted.

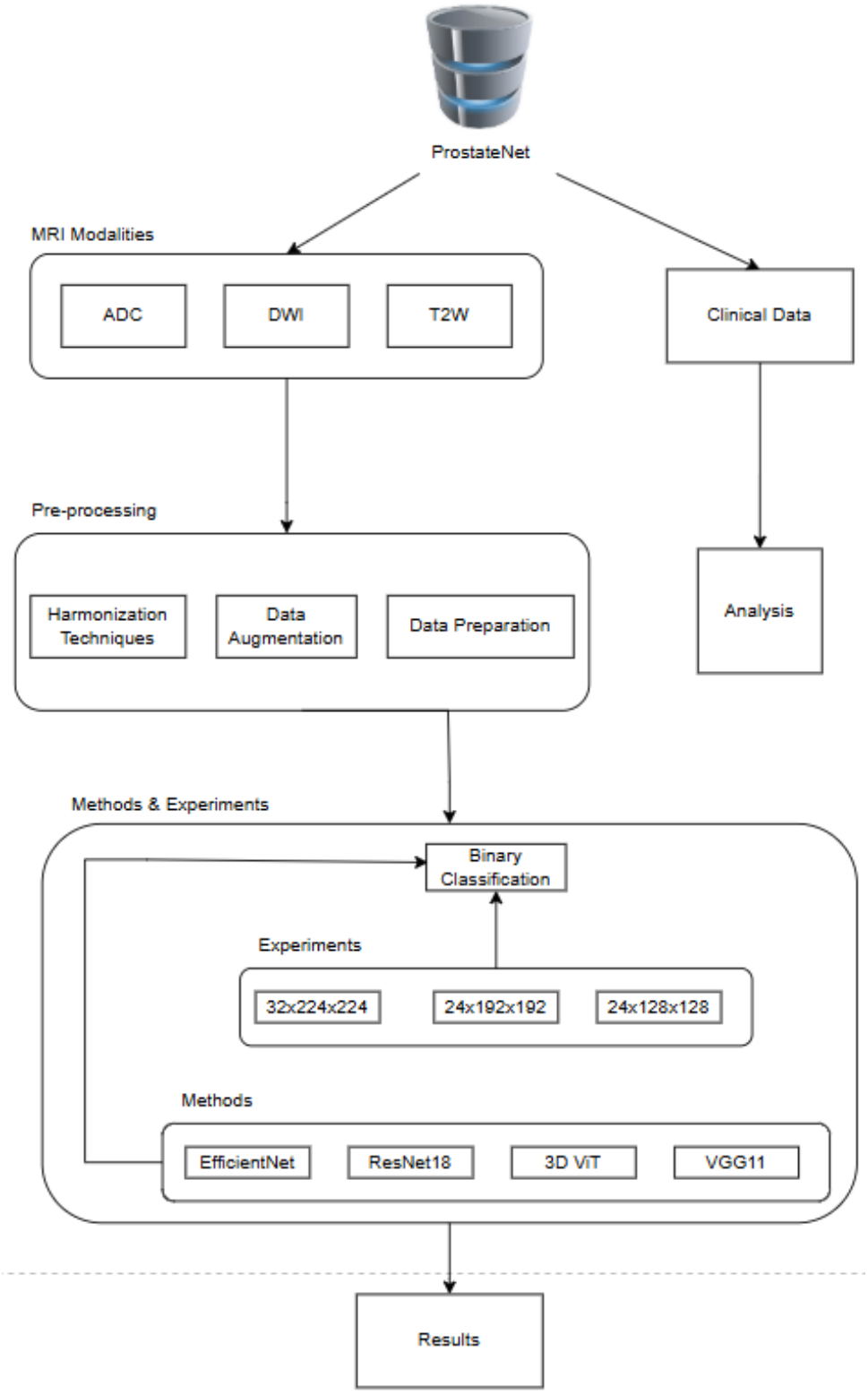


Figure 20. Overview of the proposed pipeline

3.2 ProstateNET Imaging Archive

For the first task of this thesis, which focuses on design and development of pre-processing techniques aimed at harmonizing heterogeneous PCa imaging data to improve AI model performance, we used the *ProstateNET* dataset. It's a comprehensive image and clinical data archive of anonymized prostate cancer images with over 1.5 million image representations of PCa from more than 17,000 patients. It represents the largest collection of PCa multiparametric MR (mpMRI) images including T2 weighted (T2W), Apparent Diffusion Coefficient (ADC) and Diffusion-Weighted (DWI) sequences[16]. Data was acquired by 12 clinical partners through the ProCAncer-I project[17] [74]. Siemens, Philips, Toshiba and GE scanners with variety of models and magnetic fields (1.5T and 3T) (Table 2) used by the clinical centers to obtain the MR images. Access to the ProstateNET dataset was granted by the ProCAncer-I consortium upon request, as it constitutes a private and restricted-access data resource.

Table 2. ProstateNET Scanner Models by Manufacturer

Manufacturer	*Scanner Models*
Siemens	Skyra, Aera, Prisma_fit, Verio, Avanto, MAGNETOM_ESSENZA, MAGNETON Vida, SymphonyVision, SymphonyTim, Avanto_fit, Skyra_fit, MAGNETON Sola, Spectra, Symphony, MAGNETOM Lumina, Avanto_DOT, MAGNETON Altea, Espree, TrioTim
Philips	Achieva, Ingenia, Achieva sStream, Ingenia Elition X, Ingenia Ambition X
Toshiba	Vantage Elan, Galan 3T
GE	Optima MR450w, Signa HDxt, SIGNA Artist, Optima MR360, SIGNA Explorer, SIGNA Architect, SIGNA Pioneer, SIGNA Voyager, DISCOVERY MR750w, DISCOVERY MR750, Signa Voyager, DISCOVERY MR450

[16] Source: https://prostatenet.eu/

[17] Source: https://www.procancer-i.eu/

For the purposes of this study, we used T2W, ADC and DWI images from 3935 patients followed by their corresponding clinical data. Based on the International Society of Urological Pathology (ISUP) grade groups, 1102 patients were recorded as *non-clinically significant PCa* (labelled as 0) and 2833 patients as *clinically significant PCa* (labelled as 1) for the classification task (Figure 21). The age range of the screening candidates at baseline included in the study was 40 to 91 years. Additionally, 9 patients that noticed with age value of 0 were excluded from the age analyses as these were most likely missing data.

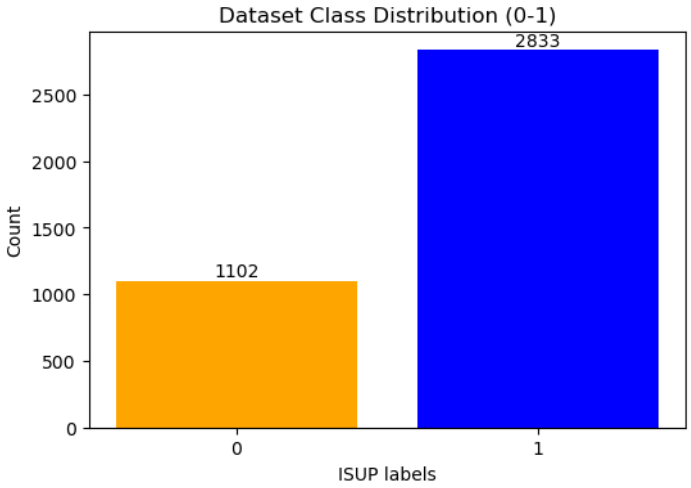


Figure 21. Dataset Class Distribution

All images were provided in Neuroimaging Informatics Technology Initiative (NIfTI) format[18] with 32x224x224 dimensions and 3x0.5x0.5 mm voxel spacing acquired from 1.5T and 3T scanners. More specifically, 528 and 574 PCa non-clinically significant patients were scanned with 1.5T and 3T MRI machines respectively. In addition, 1093 and 1740 PCa clinically significant patients were scanned with 1.5T and 3T MRI machines respectively (Figure 22). Overall, 2008 cases were acquired from Siemens, 576 from GE, 1346 from Philips and 5 from Toshiba (Figure 23).

[18] Source: https://nifti.nimh.nih.gov/

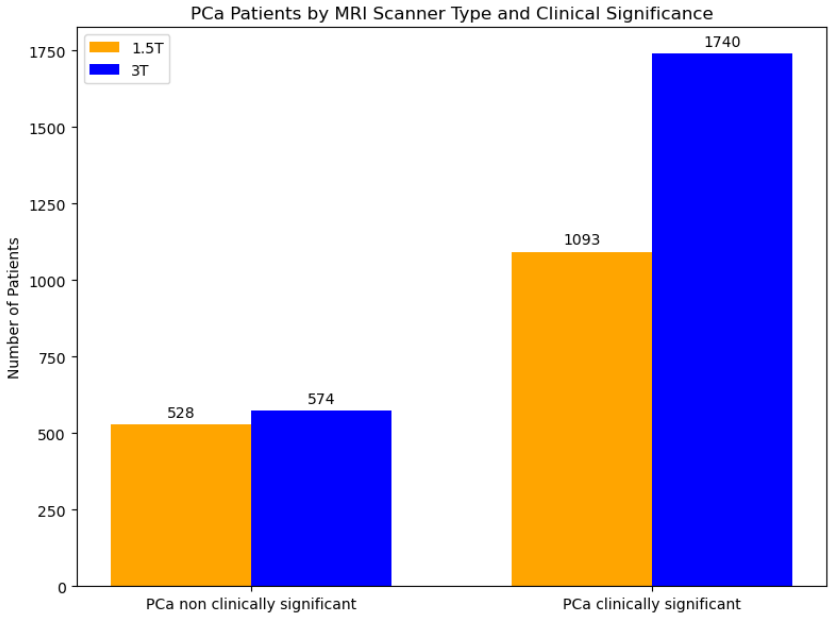


Figure 22. Distribution of PCa Patients by MRI Scanner Strength and Clinical Significance

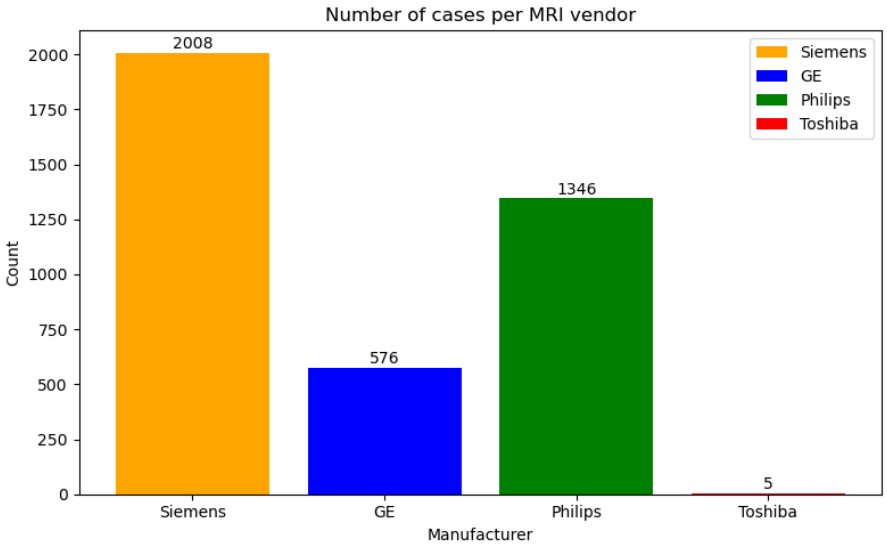


Figure 23. Number of cases per manufacturer

3.3 Pre-processing

As previously mentioned, MR images are provided in volume size of 32x224x224. However, in order to reduce the computational cost and assess if the input resolution affects models' performance, we created two new datasets with volume sizes of 24x128x128 and 24x192x192 respectively (Figure 24). Specifically, we center-cropped the images around the prostate gland using the *CropOrPad* function provided by the *TorchIO* [75] library keeping only the first 24 slices along the depth dimension as these contain the relevant anatomical information. The remaining 8 slices were excluded since they are either black slices or include non-relevant information. The harmonization techniques described in section 2.1.4 were applied consistently across all datasets.

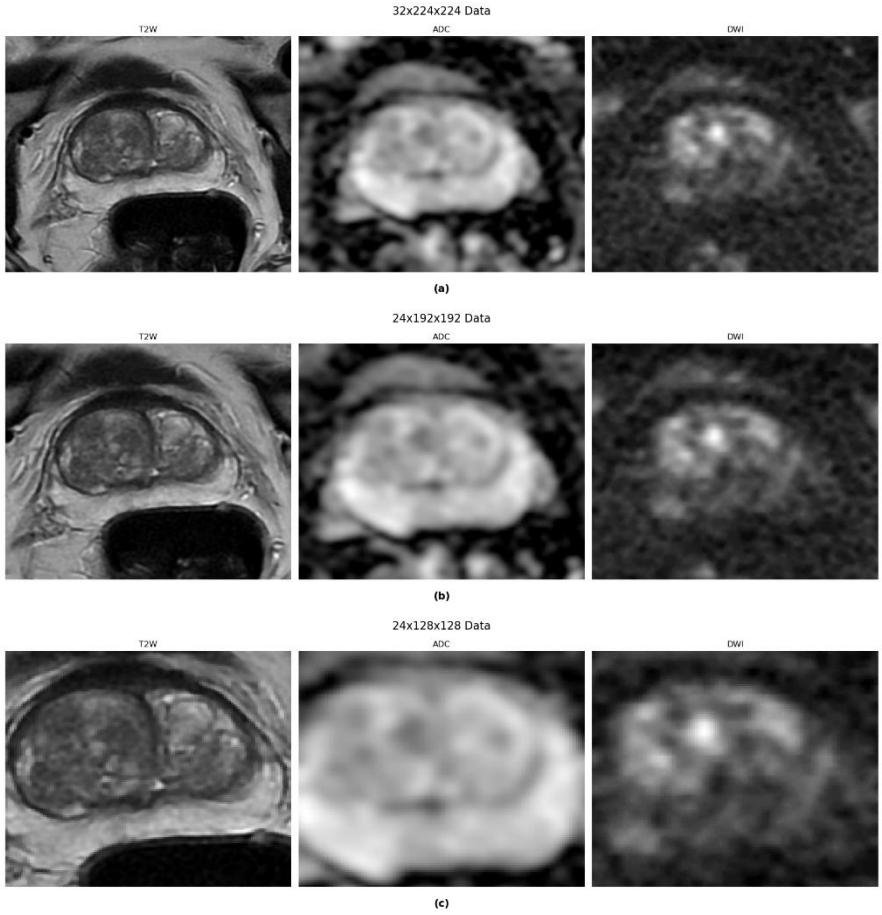


Figure 24. MRI sequences at three volume sizes (slice 16). (a) raw data 32x224x224. (b) center-crop data 24x192x192. (c) center-crop data 24x128x128

3.4 Implemented Harmonization Techniques

To address the inter-site and inter-scanner variability and increase DL models performance, we applied several harmonization techniques, including z-score, histogram, mean normalization and the combination of z-score and N4 bias field correction.

After center cropping, all the harmonization techniques were applied to all datasets and implemented using the Python programming language, employing several libraries. More specifically, TorchIO library was used to perform z-score and histogram normalization. Mean normalization was performed using *NumPy* [76], and *ANTs* (Advanced Normalization Tools) library [77] used for the N4 bias field correction. All the harmonization methods were applied to all datasets. Below, for each method, we provide intensity distribution plots between 1.5T and 3T MRI scans for each MRI sequence (ADC, DWI, T2W) before and after harmonization, from 20 randomly selected images (Figures 25-29).

Raw Data

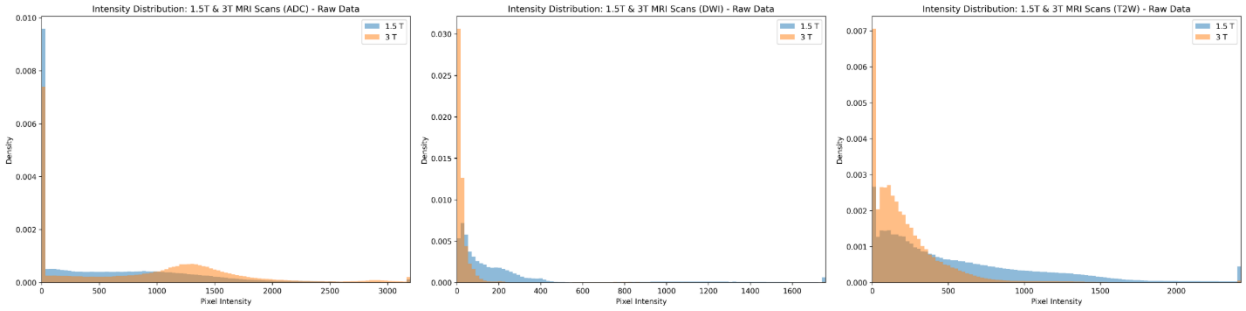


Figure 25. All MRI sequences - Raw data

The raw data demonstrates large intensity inconsistencies between 1.5T and 3T scans across all MRI sequences, indicating scanner-related variations and the need of harmonization.

Z-score Normalization

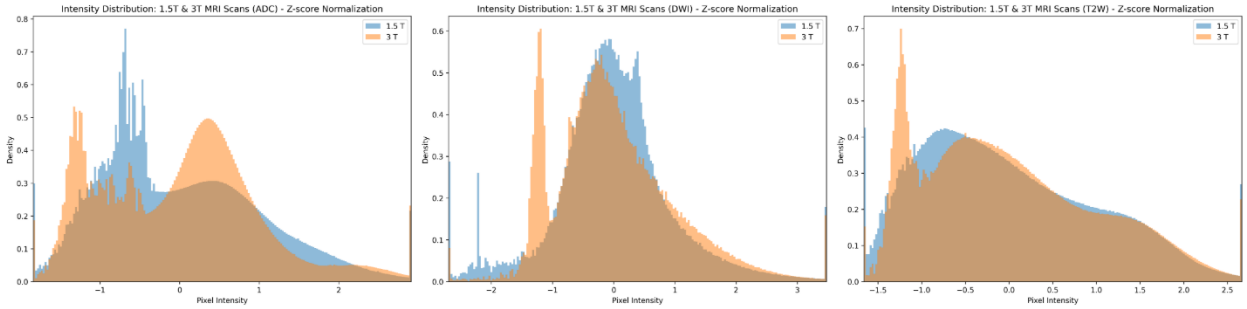


Figure 26. All MRI sequences - After z-score normalization

After applying the z-score normalization, all values are shifted to center around zero and rescaled by their standard deviation (std). We notice that using this method the intensity offset between scanners reduces, so that the overlap between 1.5T and 3T especially in DWI images becomes more noticeable.

Histogram Normalization

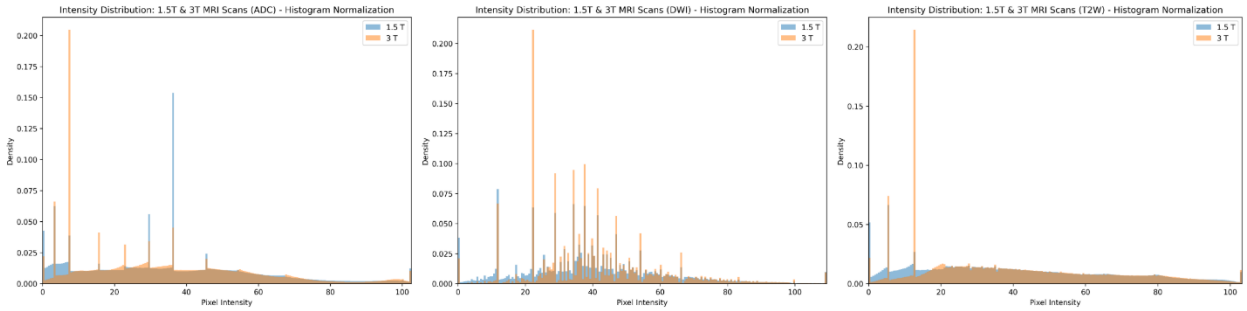


Figure 27. All MRI sequences - After histogram normalization

Using the histogram normalization technique, the intensity distribution of 1.5T and 3T scans appears more closely aligned, which is especially visible in ADC and T2W images. Specifically, for this method, a referenced histogram was created using a set of 160 images (for each MRI sequence) equally distributed across magnetic fields strengths, scanner vendors (excluding Toshiba) and PCa clinical significance. The diversity of this set likely contributed to better harmonization from varied sources effectively.

Mean Normalization

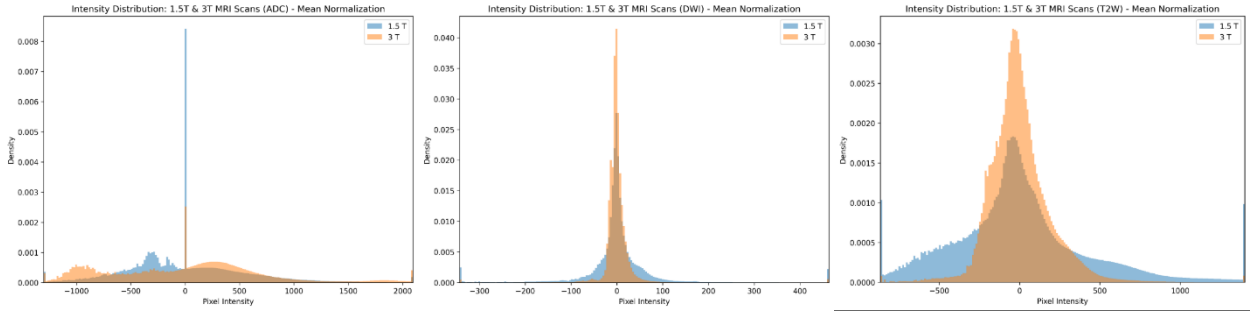


Figure 28. All MRI sequences - After mean normalization

With the application of mean normalization all intensity distributions centered around zero. However, differences in distribution shape (such as the location of peaks), remain visible between 1.5T and 3T scans. These observations suggest that the mean normalization alone is insufficient to reduce-harmonize inter-site and inter-scanner variabilities in our 3D data.

Z-score & N4 Bias Field Correction

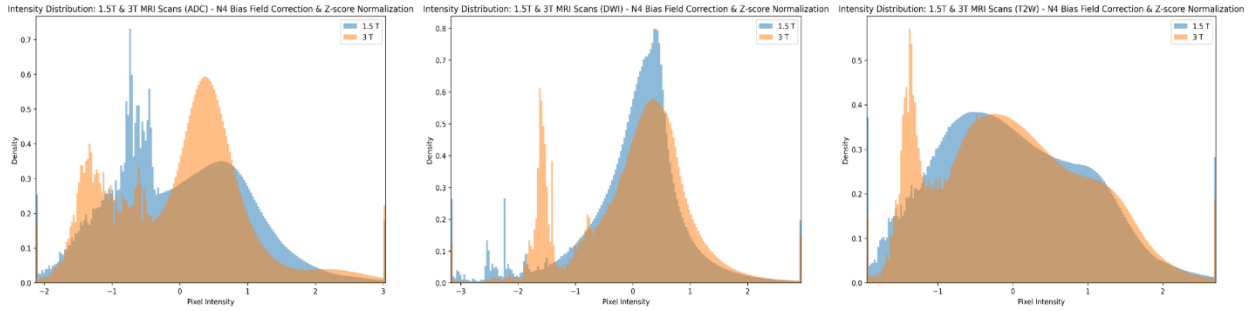


Figure 29. All MRI sequences - After z-score & N4 bias field correction

The combination of z-score normalization and N4 bias field correction showed that it improves consistency between 1.5T and 3T scans. Specifically, in DWI and T2W sequences the distributions show better overlap compared to z-score alone.

After the implementation of harmonization techniques, data were split into 70% train (2754 samples), 15% validation (590 samples) and 15% test (591 samples) sets. A stratified approach was taken to maintain balance between the classes ensuring both classes 0 (PCa non-clinically significant) and 1 (PCa clinically significant) were proportionally present in each subset as seen in Figure 30.

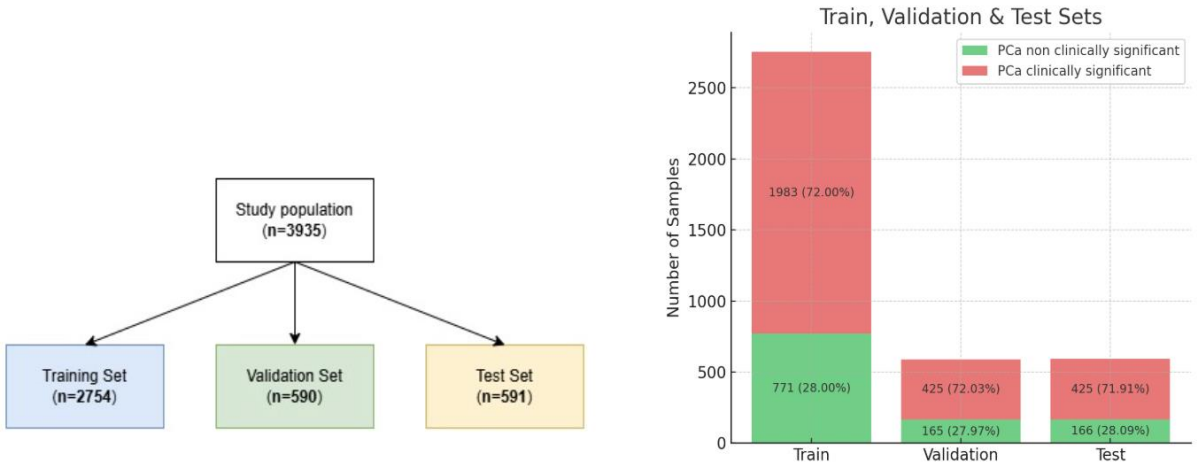


Figure 30. Study population distribution and dataset stratification

To enhance models' generalizability and robustness, we applied several data augmentation techniques during the training phase. These included random contrast adjustment, intensity and standard deviation shifting, addition of Rician noise, bias field distortion, affine transformations and random flipping. Augmentations applied to all MRI sequences in training set using the *OneOf* strategy from Medical Open Network for AI (MONAI) framework where at each iteration a single augmentation was randomly selected and applied. Specifically, MONAI framework [78] is an open-source framework for healthcare imaging which provides domain-specific tools and capabilities for AI deployment in medical applications. It is a PyTorch-based framework offering specialized algorithms, evaluation metrics and data handling, while supporting the entire AI model development lifecycle, from data preparation to deployment.

3.5 Implementation details

The process outlined in Figure 31 illustrates our deep learning approach for PCa management using MRI sequences. Following the pre-processing and data preparation stages, before model training, the MRI sequences T2W, DWI and ADC (for each patient) were stacked together into three input channels, forming the input of DL models. These input channels typically used for RGB color data in conventional image processing tasks, but here they were repurposed to represent distinct MRI sequences. To capture spatiotemporal features effectively, all DL models' conventional 2D designs were modified into 3D architectures capable of processing entire 3D MRI volumes. Specifically, EfficientNet, ResNet18, a Vision Transformer and VGG11 all in 3D

representations, employed for this task. All models implemented in Python language using the *Jupyter Notebook* environment [79] and trained separately on all datasets.

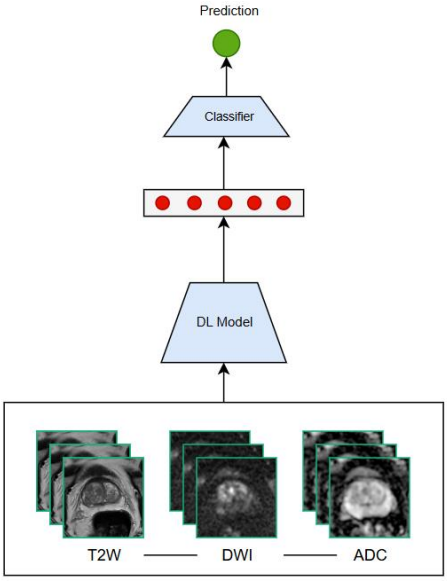


Figure 31. The 3-channel architecture

EfficientNet

The EfficientNet model, specifically the 3D Efficient-B0 architecture, re-implemented using a publicly available code repository[19] based on PyTorch deep learning framework. To adapt the model to our task, we modified several parameters such as the number of output classes (set to 1 for binary classification), input channels (set to 3 for 3D data) and dropout rate (set to 0.5 to reduce overfitting). During the training process, performance metrics are tracked to assess how effectively the model classified the input data.

The model architecture consists of 3D Mobile Inverted Bottleneck Convolutional Blocks (MBConv3D), designed to extract features from volumetric MRI scans. More specifically, in the beginning the architecture has a 3D convolution stem layer, followed by several MBConv3D blocks that increase in size as the data passes through the network. Memory-Efficient Swish activations in combination with these blocks, enhance non-linearity and preserve at the same time the gradient flow. In addition, dynamic same padding ensures compatibility with varying input

[19] Source: https://github.com/xmuyzz/3D-CNN-PyTorch/blob/master/models/EfficientNet.py

dimensions. In the end of architecture, a global Adaptive Average Pooling layer compresses the learned features and the fully connected layer produce the classification output.

ResNet18

From the ResNet architectures, we employed the 3D version of ResNet18 using the MONAI framework. Model architecture starts with a 3D convolutional layer containing 64 filters followed by batch normalization, a ReLU activation and 3D max pooling. The main core of the 3D ResNet18 consists of four sequential stages of 3D ResNet blocks and stacked Conv3D layers. The number of features channels increase from 64 to 512 while the data passes through these stages. In addition, after each convolution, batch normalization and ReLU activations applied in order to stabilize learning. An adaptive average pooling layer is used for 3D feature maps compression into a compact vector, followed by a dropout layer. Lastly, the fully connected layer produces a single output, representing the model's prediction, clinically or non-clinically significant prostate cancer.

ViT

The model architecture was re-implemented using a publicly available code repository[20], which offers several implementations of ViTs for both 2D and 3D data, using PyTorch. Specifically, we employed the 3D ViT, adapted to process our 3D MRI data, in binary classification task. Depending on the input volume size (24x128x128, 24x192x192 or 32x224x224), we adapted model's parameters such as image size and depth. Also, other parameters including the number of classes (set to 1) and the dropout (set to 0.5) modified to fit with our task.

The transformer's architecture starts with a layer called rearrangement, which flattens and reorganizes the 3D data input into a 384 tokens sequence. Each token is linearly projected into a 512-dimensional embedding space followed by a layer normalization which stabilizes and accelerates the training process. Continuing, a classify token [CLS] is added to this sequence, to pool contextual information. The model's core consists of multiple stacked transformer encoder blocks. Each one contains a multi-head self-attention mechanism and a feedforward network. In

[20] Source: https://github.com/lucidrains/vit-pytorch

each block attention, scores are computed using scaled dot-product attention with learned projections for queries, keys, and values. For regularization and stability, these layers are interleaved with layer normalization and dropout. In the feedforward network, GELU activations are used in order to introduce non-linearity, are also wrapped with dropout to prevent overfitting. The [CLS] token is extracted after the final transformer layer, passed through a final layer normalization and outputs through a fully connected linear layer to a single output corresponding to the classification task.

VGG11

From the VGG models family, we implemented the VGG11 architecture specially modified with 3D convolutions to handle 3D volumetric data.

VGG11 architecture starts with a 3D convolutional layer with four filters followed by a ReLU activation function and 3D max pooling layer to reduce the input volume dimensions. This block is followed by four blocks with convolutional layers with increased depth, 128, 256 and 512 filters. Each block consists of ReLU activations and max pooling activations. More abstract and spatial representation of the volumes is progressively extracted by these layers, where the deepest convolutional layers capture more complex anatomical patterns. After the last block of convolutional layers, there is an adaptive average pooling layer that reduces the feature maps into 1x1x1 spatial size. These maps are then flattened and pass through a fully connected classifier. The classifier consists of two dense layers, with 4096 units each, together with ReLU activations and dropout. Lastly, a linear output layer produces the binary classification output.

Experimental Settings

All models used in this task were trained using the Adam optimizer with a learning rate, L1 and L2 regulation strategies of 3×10^{-5}. A weighted binary cross-entropy was used as the loss function. Data augmentation was employed including appearance augmentations (bias, contrast, noise, intensity), flipping and rotation.

The experiments were carried out on a machine equipped with an Intel® Xeon® @ 2.40GHz processor, 2x NVIDIA GeForce GTX 1080 Ti (each with 16GB VRAM) and 32GB DDR4 RAM. PyTorch 2.6.0 was used during the experiments and executed in Python 3.12.7.

3.6 Results

In this section we present the results of the experiments conducted using four DL models especially modified to handle 3D data EfficientNet, ResNet18, ViT and VGG11 in PCa clinical significance binary classification task. The models were trained in all volume sizes MRI data (32x224x224, 24x192x192, 24x128x128) employing all harmonization techniques as described in section 2.1.4 in each dataset. The goal of this work was to enhance models' performance using harmonization techniques applied to multi-vendor and multi-centric MR images. We conducted experiments based on three different data inputs: input case 1: 24x128x128 center-cropped data, input case 2: 24x192x192 center-cropped data, input case 3: 32x224x224 original data. In each input case, we trained all models using raw data and each of the harmonization techniques. Finally, we compared the models' performance between raw and harmonized data to assess the impact of harmonization methods on enhancing their capability to process heterogeneous imaging data effectively. All outcomes for each input case are presented in Tables 3-5.

Input case 1: 24x128x128 Center-cropped Data Experiments

In this dataset, VGG-11 model achieved the highest AUC score of 64.28% and balanced accuracy of 59.77% on the **raw data** compared to other models as seen in Table 3. The worst performance was achieved by EfficientNet with AUC of 56.42%. However, despite the lowest AUC score, EfficientNet achieved the highest specificity among all models, 55.42%. For **z-score normalization** alone and in **combination with N4 bias field correction**, AUC score of ResNet18 model increased by 7.06% (from 63.70% to 70.76%) and 0.28% (from 63.70% to 63.98%) respectively. Also, the highest specificity of 74.08% using **z-score normalization** method achieved by EfficientNet and the highest balanced accuracy of 65.03% by ResNet18 compared to raw data. For the **N4 and z-score combination**, ResNet18 showed the highest specificity of 53.01%. Using the **histogram normalization** method, three models including EfficientNet, ResNet18 and VGG11 showed increased performance. Specifically, for EfficientNet the AUC score increased by 6.27% (from 56.42% to 62.69%) and for ResNet18 and VGG11 increased by 4.58% (from 63.70% to 68.28%) and 1.54% (from 64.28% to 65.82%) respectively. Moreover, EfficientNet and VGG11 showed the highest specificity and balanced accuracy, 82.53% and 61.94% respectively. Lastly, for **mean normalization** ResNet18 outperformed the other models with 1.06% increased AUC (from 63.70% to 64.76%) compared to its performance in raw data.

Best specificity and balanced accuracy score achieved by VGG11 and ResNet18, 51.81% and 59.13% respectively.

Table 3. Classification results - 24x128x128 data

Harmonization Method	*Model*	*Specificity*	*Balanced Accuracy*	*AUC*
Raw Data	EfficientNet	55.42%	55.71%	56.42%
	ResNet18	39.80%	58.23%	63.70%
	ViT	36.80%	59.20%	61.49%
	VGG11	35.54%	59.77%	64.28%
Z-score	EfficientNet	**74.09%**	49.99%	50.24%
	ResNet18	**67.46%**	**65.03%**	**70.76%**
	ViT	**53.61%**	56.57%	61.44%
	VGG11	0%	50.00%	50.00%
N4+Z-score	EfficientNet	11.44%	51.61%	56.42%
	ResNet18	**53.01%**	57.68%	**63.98%**
	ViT	**49.39%**	54.58%	58.00%
	VGG11	0%	50.00%	50.00%
Histogram Normalization	EfficientNet	**82.53%**	**55.97%**	**62.69%**
	ResNet18	**56.63%**	**61.84%**	**68.28%**
	ViT	**55.42%**	58.18%	61.05%
	VGG11	**59.64%**	**61.94%**	**65.82%**
Mean Normalization	EfficientNet	44.58%	54.64%	55.07%
	ResNet18	36.14%	**59.13%**	**64.76%**
	ViT	34.94%	56.41%	60.30%
	VGG11	**51.81%**	59.08%	**66.17%**

Input case 2: 24x192x192 Center-cropped Data Experiments

Regarding the 24x192x192 dataset, VGG11 achieved the highest performance outperforming all other models with AUC score of 64.37% and balanced accuracy of 62.14% on test set using **raw data** as seen in Table 4. On the other hand, EfficientNet showed the worst performance by achieving AUC score of 55.93%. The highest specificity was reached by ResNet18 correctly classifying 66.86% of the minority class. For **z-score normalization** EfficientNet, ResNet18 and ViT showed increased performance compared to raw data, AUC increased by 0.16% (from 55.93% to 56.09%), 6.15% (from 63.52% to 69.67%) and 0.71% (from 60.25% to 60.96%) respectively. The best specificity (83.13%) achieved by EfficientNet and the highest balanced accuracy (64.73%) by ResNet18. EfficientNet, ResNet18 and VGG11 increased their performance compared to raw data using the **histogram normalization** method. Specifically, EfficientNet increased AUC by 4.74% (from 55.93% to 60.67%), ResNet18 by 7.76% (from 63.52% to 71.28%) and ViT by 1.9% (from 60.25% to 62.15%). EfficientNet and ResNet18 increased their specificity and balanced accuracy, 65.06% and 63.72% respectively. Continuing, using the **mean normalization** technique, ResNet18 AUC increased by 1.03% (from 63.52% to 64.55%) and ViT by 1.18% (from 60.25% to 61.43%). Also, the ViT model showed the best balanced accuracy of 58.40%. In contrast, the combination of **N4 bias correction and z-score normalization** didn't show any performance improvement of AUC score for any of the models compared to their performance using raw data. Only specificity and balanced accuracy increased compared to raw data, 74.10% and 58.36% respectively.

Table 4. Classification results - 24x192x192 data

Harmonization Method	*Model*	*Specificity*	*Balanced Accuracy*	*AUC*
Raw Data	EfficientNet	37.95%	55.65%	55.93%
	ResNet18	66.86%	61.46%	63.52%
	ViT	53.01%	57.96%	60.25%
	VGG11	36.75%	62.14%	64.37%
Z-Score	EfficientNet	**83.13%**	53.33%	**56.09%**
	ResNet18	**66.87%**	**64.73%**	**69.67%**
	ViT	53.01%	54.39%	**60.96%**
	VGG11	0%	50.00%	50.00%
N4 + Z-Score	EfficientNet	**74.10%**	46.22%	47.68%
	ResNet18	63.25%	59.86%	61.73%
	ViT	50.60%	**58.36%**	59.14%
	VGG11	0%	50.00%	50.00%
Histogram Normalization	EfficientNet	**65.06%**	54.88%	**60.67%**
	ResNet18	66.27%	**63.72%**	**71.28%**
	ViT	39.76%	57.41%	**62.15%**
	VGG11	0%	50.00%	50.00%
Mean Normalization	EfficientNet	36.14%	55.37%	59.23%
	ResNet18	56.02%	60.72%	**64.55%**
	ViT	39.16%	**58.40%**	**61.43%**
	VGG11	32.53%	61.09%	64.21%

Input case 3: 32x224x224 Original Data Experiments

Using the original dataset, ResNet18 achieved the best AUC score of 64.05% in **raw data**, outperforming all the other models. The highest balanced accuracy was achieved by VGG-11, 60.76%. Nevertheless, the ViT model had the worst performance with an AUC score of 60.46% as illustrated in Table 5. The highest specificity was reached by EfficientNet correctly classifying 63.85% of the minority class. For **z-score normalization** ResNet18 and ViT increased their performance compared to raw data by 3.22% (from 64.05% to 67.27%) and 2.9% (from 60.46% to 63.36%) respectively. Also, the ResNet18 model showed the highest balanced accuracy and specificity, 68.67% and 61.51% respectively. EfficientNet, ResNet18 and ViT models showed increased performance when trained on MR images harmonized using the **combination of N4 and z-score normalization**. Specifically, AUC score increased by 0.62% (from 61.58% to 62.20%), 0.64% (from 64.05% to 64.69%) and 1.43% (from 60.46% to 61.89%) respectively. Models ResNet18 and ViT increased their specificity to 63.86% and 53.61% respectively, while EfficientNet showed increased balanced accuracy of 57.24%. In addition, only ResNet18 and ViT models came up with better performance using **histogram normalization**. In detail, ResNet18 increased its AUC score by 5.77% (from 64.05% to 69.77%) and ViT by 3.37% (from 60.46% to 63.83%) compared to their raw data performance. EfficientNet and ResNet18 achieved the highest specificity (83.73%) and balanced accuracy (62.04%) respectively. Lastly, **mean normalization** technique didn't show any improvement in models' performance compared to raw data in AUC score and balanced accuracy. Only ViT showed improvement in specificity (43.37%).

Table 5. Classification results - 32x224x224 data

Harmonization Method	*Model*	*Specificity*	*Balanced Accuracy*	*AUC*
Raw Data	EfficientNet	63.85%	55.57%	61.58%
	ResNet18	32.53%	60.27%	64.05%
	ViT	34.34%	57.52%	60.46%
	VGG11	34.93%	60.76%	63.27%
Z-Score	EfficientNet	25.90%	**58.48%**	63.03%
	ResNet18	**68.67%**	**61.51%**	**67.27%**
	ViT	**45.78%**	**59.48%**	**63.36%**
	VGG11	0%	50.00%	50.00%
N4+Z-score	EfficientNet	40.36%	**57.24%**	**62.20%**
	ResNet18	**63.86%**	59.22%	**64.69%**
	ViT	**53.61%**	56.81%	**61.89%**
	VGG11	0%	50.00%	50.00%
Histogram Normalization	EfficientNet	**83.73%**	52.69%	57.54%
	ResNet18	**73.49%**	**62.04%**	**69.77%**
	ViT	**54.82%**	**60.12%**	**63.83%**
	VGG11	0%	50.00%	50.00%
Mean Normalization	EfficientNet	25.30%	53.59%	54.72%
	ResNet18	**42.17%**	57.44%	62.34%
	ViT	**43.37%**	57.45%	60.44%
	VGG11	0%	49.76%	58.28%

3.7 Discussion

In the overall assessment of model performance across all experimental settings, all 3D architectures showed promising performance in the binary classification task, distinguishing between clinically and non-clinically significant prostate cancer using multi-centric and multi-vendor MRI data. Among all the experimental settings, ResNet18 achieved the highest AUC score and balanced accuracy, especially when the harmonization techniques histogram and z-score normalization applied. For instance, in the original volume size 32x224x224 using histogram normalization, ResNet18 achieved an AUC and balanced accuracy of 69.77% and 62.04% respectively, outperforming all the other models. Similarly, the highest specificity achieved by EfficientNet (83.73%) using the same harmonization method demonstrating high capability in minority class (non-clinically significant PCa) identification.

Across all the experiments, indeed, harmonization techniques increased models' performance, with histogram and z-score normalization offering the best improvements across all the DL models and volume sizes. Our findings confirmed the effectiveness of harmonization techniques in reducing domain shift and enhancing generalizability in multi-centric and multi-vendor studies. However, the harmonization technique N4 bias field correction, in combination with z-score normalization showed inconsistent results and sometimes poor performance. Generally, our results highlight the importance of harmonization techniques in the pre-processing pipelines and demonstrate the potential of 3D DL models, when data are properly harmonized, to handle effectively complex imbalanced PCa data.

Impact of Different Volume Sizes

We noticed that smaller volume size MR images overall led to lower AUC scores and balanced accuracies, likely due to reduced spatial information. Tumor-related features could be lost during the center-cropping, limiting in that way the model's ability to distinguish between two classes. In contrast, larger volumes such as the 32x224x224 preserved more anatomical features related to prostate, allowing DL models such as ResNet18 and ViT to extract more meaningful features, as shown by their increased AUC and balanced accuracy scores. Nevertheless, we support that despite their advantages, larger volume sizes can introduce model complexity and increase memory demands, hindering harmonization benefits, especially when simple model architectures like VGG11 are applied. Specifically, in our results, VGG11 consistently failed to gain meaningful

information when harmonized 32x224x224 data used, resulting in minimal or no improvement in AUC and balanced accuracy.

Effect of Harmonization Techniques

Histogram normalization improved model performance especially for ReNet18 and ViT models, across all volume sizes and it showed that can reduce inter-center variability in image intensity distribution.

Z-score normalization alone showed improvements in ResNet18 model by improving the AUC score by 7.06% in 24x124x124 volume size data suggesting that standardized pixel intensity benefits this architecture. However, in **combination with N4 bias field correction** it showed poor performance using the VGG11 model (50% AUC across all settings) indicating that VGG11 is very sensitive in certain pre-processing methods. Although this combination has been previously reported in the literature as a harmonization technique in multi-centric studies, our experiments showed limited performance improvement. We support that this method is either no suitable for our 3D PCa data or the combination of two normalization techniques negatively affected anatomical and meaningful patterns that models learn from.

In addition, **mean normalization** had limited impact across all model architectures and volume sizes. We support that this harmonization technique was insufficient to reduce the inter-site heterogeneity in our PCa MR images, likely due to its simplistic approach.

Models' Performance in 3D PCa Imbalanced Data

As we showed in the previous section, (Section 3.2) the dataset we used was imbalanced (2833 PCa clinically significant (class 1) and 1102 PCa non-clinically significant patients (class 0), which affected specificity across our experiments. Most of the models achieved higher sensitivity (class 1) while showing reduced specificity (class 0). This is easily noticeable in raw data, where specificity scores were often < 40% indicating bias toward the majority class. Harmonization techniques such as histogram and z-score normalization increased models' performance (especially ResNet18 and EfficientNet), achieving higher specificity and more balanced performance. More specifically, **ResNet18** demonstrated the most consistent performance improvement across all volume sizes and harmonization techniques likely due to each residual architecture, allowing better generalization. **EfficientNet** showed inconsistent results. In

several harmonization techniques it achieved high specificity (e.g. z-score 74.09%) but often decreased AUC and balanced accuracy scores, indicating overfitting. Continuing, **ViT** model in general underperformed compared to other CNN models, particularly on raw data. Overall, transformers need more data compared to CNNs in order to achieve high performance. In our case, the datasets that we used were limited for a transformer architecture. This explains why ViT did not consistently improve its performance using harmonization techniques and showed only few improvements across different settings. Lastly, **VGG11** model seemed to be very sensitive to harmonization methods with poor performance in z-score normalization alone and in combination with the N4 bias field correction (AUC score 50%).

4. Cardiotoxicity risk prediction

4.1 Introduction

In this section we provide a detailed description of the implementation process for the second task of this work, cardiotoxicity risk prediction, as illustrated in Figure 32. Heart ultrasound video sequences along with their clinical metadata from breast cancer patients, provided by the CARDIOCARE EU project, were used in this task. Two CNN architectures, 3D ResNet18 and I3D trained in two scenarios, 1) using only video sequences and 2) combining video sequences with tabular (clinical) data. Moreover, the EchoNet pre-trained model was fine-tuned on video sequences only data. Prior to model training, pre-processing techniques such as data normalization, cleaning, resizing and cropping performed.

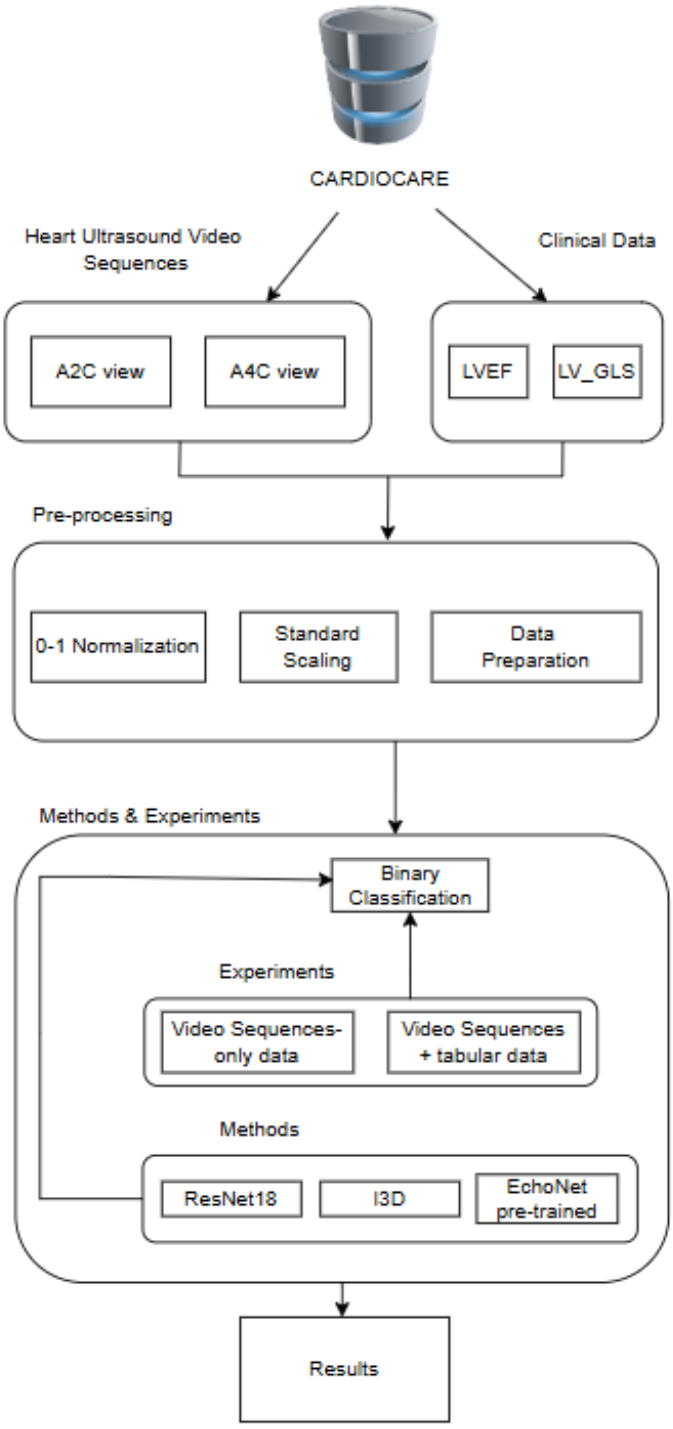


Figure 32. Overview of the proposed pipeline in cardiotoxicity risk prediction

4.2 CARDIOCARE dataset

In the second part of this thesis, we used baseline 3D echocardiography video sequences (ultrasounds) data from elderly breast cancer patients, acquired before the start of chemotherapy treatment, as part of the CARDIOCARE EU-funded project[21]. Specifically, multidimensional data such as imaging (echocardiography and mammography), genomic, biochemical and behavioral were collected. The goal of this project is to improve the quality of life and the overall care which is provided to breast cancer patients above 65 years of age[22]. Access to the CARDIOCARE dataset was granted upon request, as it constitutes a private and restricted-access data resource.

For the purposes of this study, we used 497 apical-4-chamber (A4C) and 373 apical-2-chamber (A2C) ultrasound video sequences from breast cancer patients followed by their clinical data. From clinical data, we used the Left Ventricular Ejection Fraction (LVEF) and Left Ventricular Global Longitudinal Strain (LV_GLS) measurements. Due to data anonymization protocols employed during data collection, the age range of the patients could not be calculated. Each patient was manually classified as cardiotoxic (labelled as 1) or non-cardiotoxic (labelled as 0) based on clinical assessments. More specifically, in A4C view data, 139 samples were recorded as cardiotoxic and 117 as non-cardiotoxic. However, 241 samples were excluded from the dataset because their cardiotoxicity labels were missing (recorded as NaN), making them unusable in classification tasks. Similarly, in A2C view data, 112 samples were labeled as cardiotoxic and 89 as non-cardiotoxic while the remaining 172 samples were excluded from the dataset due to missing cardiotoxicity label (NaN) (Figure 33).

[21] Source: https://cardiocare-project.eu/

[22] Source: https://cardiocare-project.eu/wp-content/uploads/2024/07/The-CARDIOCARE-project-Report.pdf

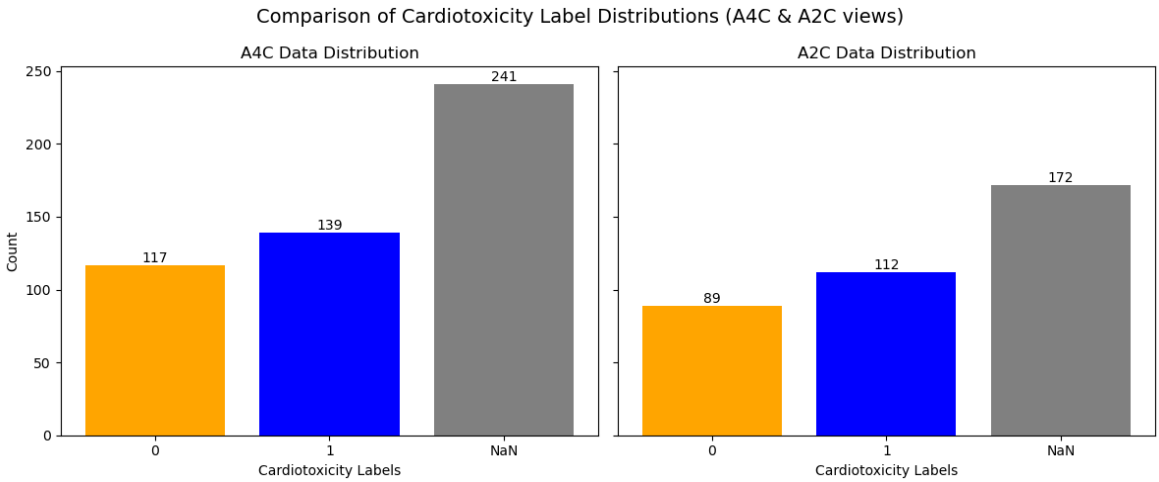


Figure 33. Cardiotoxicity dataset class distribution

All video sequences were provided in Digital Imaging and Communications in Medicine (DICOM)[23] format with a variety of dimensions including 1016x708, 800x600, 1024x576 and 1137x916. The number of frames per video sequence varied, with the maximum of 483 and the minimum 12 frames for A4C view data. Additionally, for A2C view data the maximum number of frames was 515 and the minimum 24 frames. All echocardiography data were acquired by GE and Philips systems, with the overwhelming majority of ultrasounds performed on GE devices. Moreover, the overall quality of the ultrasounds was assessed and characterized as *good*, *medium* and *poor*. Specifically, for the A4C view, 41.6% of the recordings were rated as good, 32.8% as medium and the rest 25.6% as poor. In addition, for the A2C view 35.1% of the recordings were rated as good, 38.1% as medium and the rest 26.8% as poor. The distribution is shown in Figure 34.

[23] Source: https://www.dicomstandard.org/

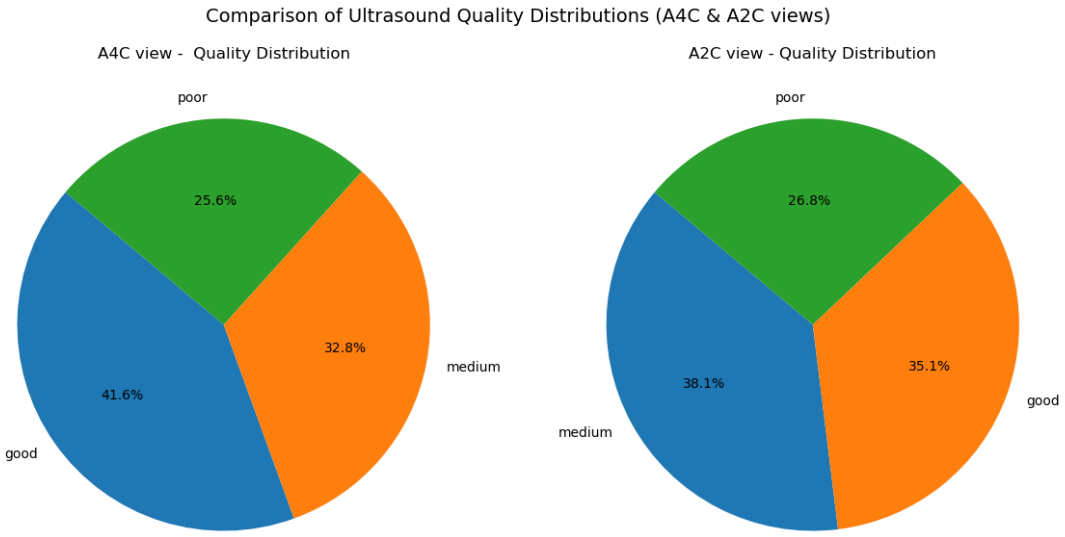


Figure 34. Ultrasounds overall quality

4.3 Pre-processing

As mentioned in the dataset description, video sequences varied in dimensions and simultaneously the original resolutions were computationally heavy for training (e.g. 1016x708). To address this, we resized all video sequences using the *Lanczos* resampling method as implemented in the *PIL* library[24]. Frames were resized to 112x112 pixel size (based on the EchoNet-Dynamic study data size) for the ResNet18 and EchoNet pre-trained model to ensure comparable results and to 224x224 for the I3D model. In addition, to focus only on the heart region, each video was cropped and masked removing irrelevant text and information outside of the scanning sector, heavily based on the EchoNet-Dynamic dataset pre-processing method (section 2.3.3). Lastly, all frames were normalized using 0-1 normalization to standardize pixel intensities across samples. Clinical measurements such as LV_GLS and LVEF were standardized using the *StandardScaler* function as provided by the *scikit-learn* library [80]. Missing values in these variables were replaced with the mean of the respective columns. Furthermore, the number of frames per sequence also varied. To manage this, we identified the number of frames that appear most frequently, which was found to be 34. Sequences exceeding this number were trimmed, while

[24] Source: https://github.com/python-pillow/Pillow

the shorter ones were padded with black frames (zero-valued) in order to ensure the same input dimensions. The figure below (Figure 35) illustrates the aforementioned method using a 28-frames ultrasound video.

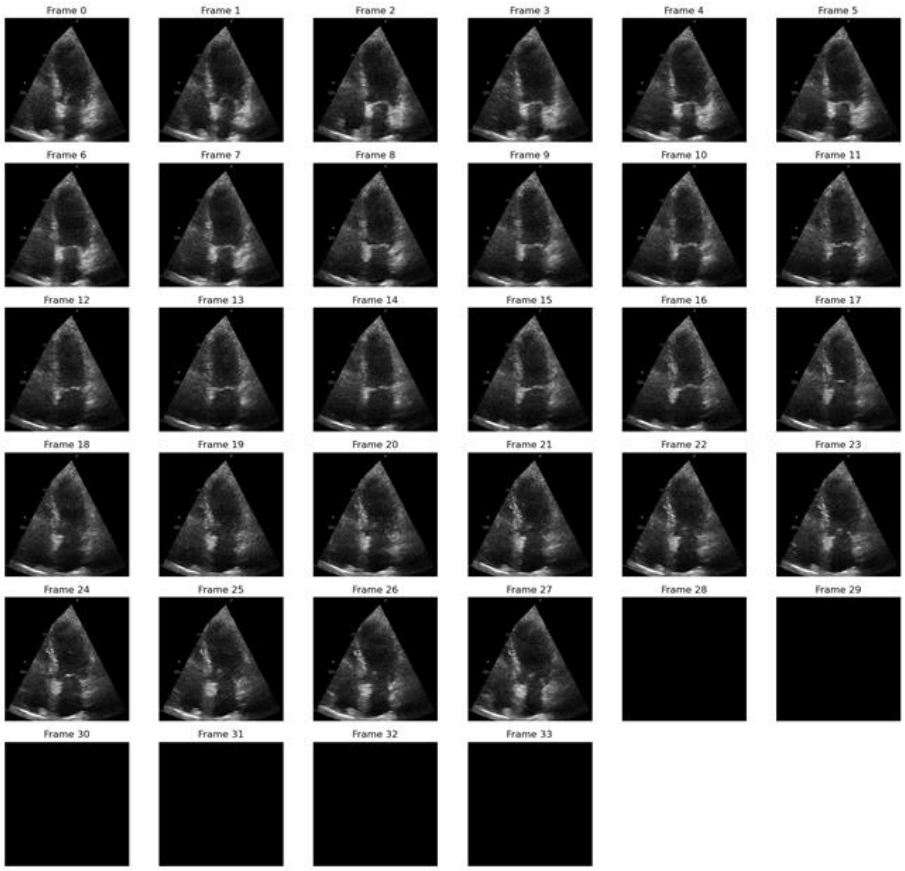


Figure 35. A 28-frame video sequence padded with zero-valued frames to reach the most frequent number of frames (34)

Data from both A2C and A4C views were split into 70% train, 15% validation and 15% test sets. For the A4C view (201 samples in total), 179 samples were allocated to training set, 38 to validation set and 39 to test set. Similarly, for the A2C view (256 samples in total), data splitted into 140 training samples, 30 validation samples and 31 test samples. A stratified sampling approach was taken to maintain balance between the classes ensuring both cardiotoxic and non-cardiotoxic classes were proportionally present in each subset as illustrated in Figures 36-37.

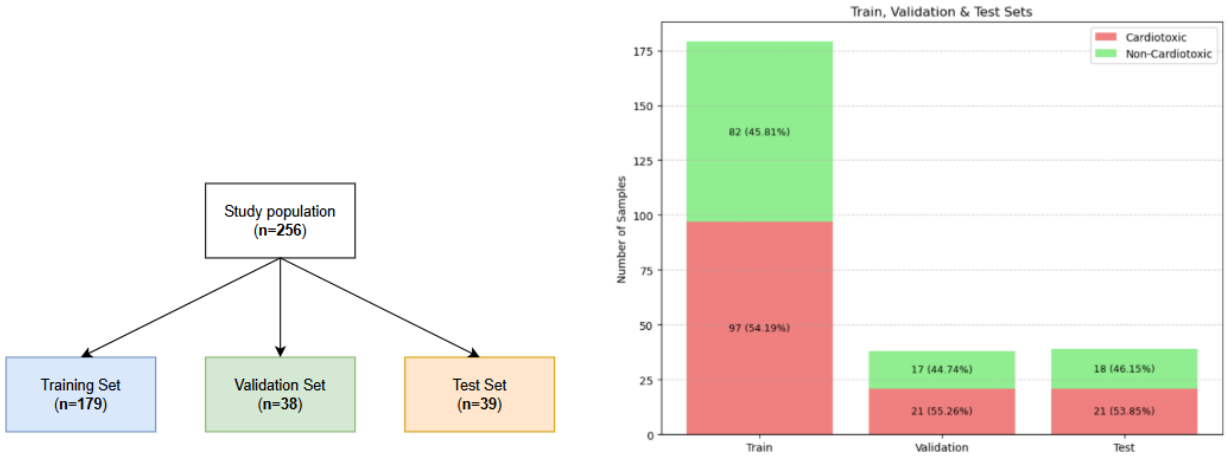


Figure 36. A4C view study population distribution and dataset stratification

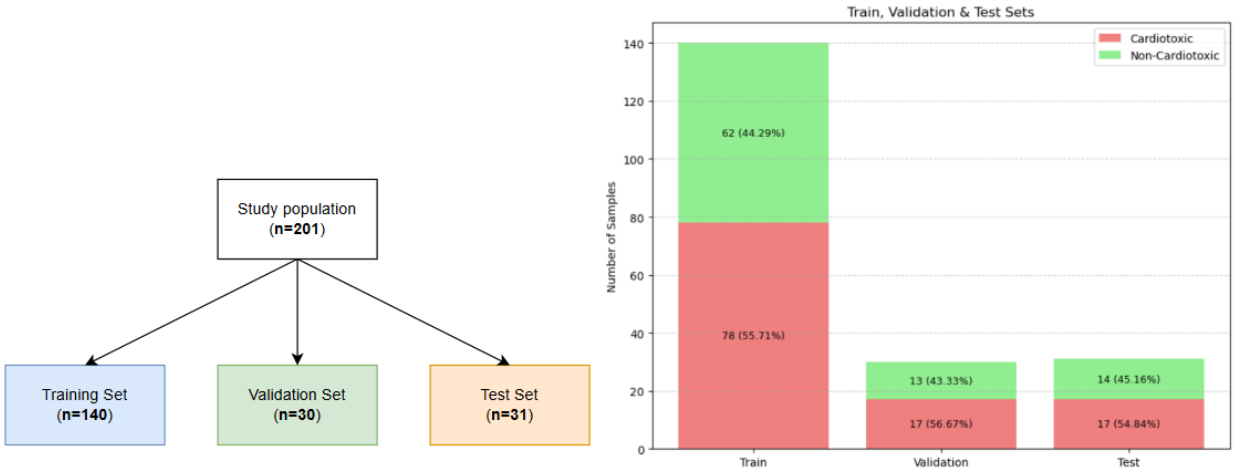


Figure 37. A2C view study population distribution and dataset stratification

4.4 Implementation details

After the pre-processing phase, data were ready to be fed into the CNN models (ResNet18, I3D, EchoNet pre-trained), for training. Instead of using ResNet18's conventional 2D architecture, we modified the model to a 3D architecture capable of capturing both spatial and temporal patterns present in the ultrasound sequences. Each model was trained under two scenarios: 1) Using only the ultrasound video data and 2) using ultrasound video data followed by their corresponding tabular echocardiographic parameters such as LV_GLS and LVEF values. Specifically, in the second scenario, the extracted image features were concatenated with the tabular features, allowing

the models to integrate both visual and clinical information in their predictions. EchoNet pre-trained trained only in video data. Deep learning models' pipelines for both scenarios are illustrated in Figure 38. Also in this task, all models implemented in Python language using the Jupyter Notebook environment where each model trained five times with different random seed in data splitting.

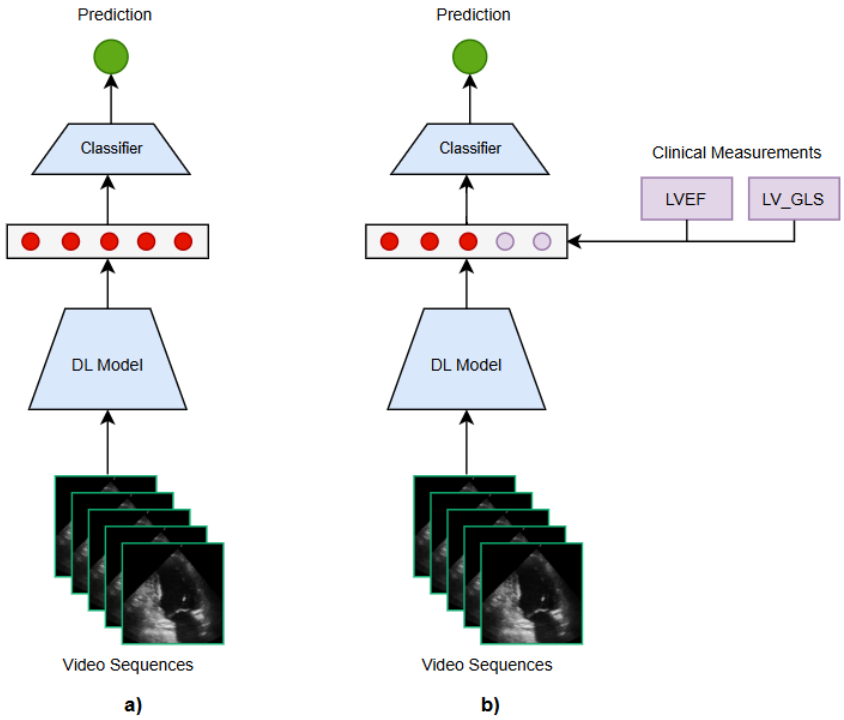


Figure 38. a) Model using only ultrasound video data, b) Model using both ultrasound video data and tabular clinical measurements (LV_GLS and LVEF)

ResNet18

For this task, the ResNet18 model was implemented in the same way as described in section 3.5. In the case of the ultrasound sequences-only scenario, the 3D ResNet18 backbone extracted a 512-dimensional spatiotemporal feature vector output which was passed to the fully connected layer, through a dropout layer, to produce the final prediction. Similarly for the second scenario, with the combination of ultrasound sequences with their corresponding tabular data (LVEF, LV_GLS) the same backbone was used for video feature extraction. In addition, the clinical parameters were passed through a small multilayer perceptron (MLP) and converted into 32-dimensional feature vector. This tabular feature vector concatenated with the 512-dimensional image feature vector,

forming a 544-dimensional feature representation that passed to the fully connected layer for the final classification.

I3D (Inception-V1)

From the Inception architectures, we employed the I3D version of Inception-V1 as introduced in Carreira and Zisserman study [12]. In our study, we used only the RGB stream of I3D that takes video frames as input. Specifically, the model inflates 2D ConvNet filters and pooling kernels into 3D, enabling spatiotemporal feature learning from video sequences. The implementation is based on PyTorch code provided by Piergiacomi in GitHub[25].

Model's architecture begins with a 3D convolution layer containing 64 filters of 7x7x7 size, followed by batch normalization, a ReLU activation function and a 3D max pooling layer. The main body of the network consists of multiple inception modules, mainly adapted from the 2D Inception-V1 architecture, to process 3D feature maps. These modules are stacked sequentially allowing the model to learn temporal dynamics alongside spatial patterns, with the number of feature channels increasing progressively with depth.

In the original paper [12] model was trained on 64-frame video sequences sampled at 25 fps. In our implementation, we used 34-frame video sequences since the number of frames that appear most frequently is 34. Moreover, after the convolutional and inception layers there is a global average pooling layer that aggregates spatiotemporal features, followed by a 1x1x1 convolutional layer that outputs the final prediction, cardiotoxicity or no-cardiotoxicity.

As with the ResNet18 backbone, the I3D model was also employed in the same two scenarios. The rest of the pipeline remained the same, only the feature extraction backbone replaced by the I3D architecture.

Experimental Settings

All models used in this task were trained using Adam optimizer with a learning rate, L1 and L2 regulation strategies of 3×10^{-5}. A weighted binary cross-entropy was used as the loss function. The experiments were carried out on a NVIDIA workstation equipped with an AMD EPYC 7742 64-Core processor, 4x NVIDIA A100 (each with 40GB VRAM) and 514GB RAM. PyTorch 2.7.0 was used during the experiments and executed in Python 3.10.12.

[25] Source: github.com/piergiaj/pytorch-i3d

4.5 Results

In this section we present the results of the experiments conducted using three DL models especially modified to handle heart ultrasound video sequences, 3D ResNet18, I3D, EchoNet pre-trained for cardiotoxicity risk prediction (binary classification task). The models were trained in two scenarios: 1) using ultrasound video sequences-only data and 2) the combination of ultrasound video sequences with their corresponding tabular data (LVEF and LV_GLS measurements) in two spatial sizes 112x112 and 224x224. The goal of this work was to investigate how 3D AI models can assist in cardiotoxicity risk prediction using full 3D heart ultrasound video sequences. We conducted two categories of experiments: 1) ResNet18 and I3D models trained at 224x224 spatial size and 2) ResNet18 and EchoNet pre-trained trained at 112x112 (I3D trained only at 224x224). In each category, we trained all models (except EchoNet pre-trained) using A2C and A4C views five times with different random seeds. EchoNet pre-trained trained only in 112x112 A4C view using only video data. Finally, we compared their performance using the mean AUC in both scenarios and categories of experiments to assess the impact of 3D video sequences on cardiotoxicity risk prediction. All outcomes for each category and experiment are presented in Tables 6-9.

A2C View

Table 6. A2C view - Video sequences only data

Model	***Frames Size***	***Mean AUC***
ResNet18	112x112	55.80%
ResNet18	224x224	**56.81%**
I3D	224x224	46.26%

Table 7. A2C view - Video sequences + Tabular Data

Model	***Frames Size***	***Mean AUC***
ResNet18	112x112	**56.05%**
ResNet18	224x224	50.34%
I3D	224x224	55.12%

In the **A2C** view, ResNet18 model achieved higher performance compared to I3D using video sequences-only data, with mean AUC score of 55.80% at 112x112 and 56.81% at 224x224, compared to 46.26% of I3D architecture. With the addition of tabular data ResNet18's performance at 224x224 decreased 6.47% (from 56.81% to 50.34%) and slightly improved at 112x112 (from 55.80% to 56.05%) while I3D's performance improved 8.86% (from 46.26% to 55.12%).

A4C View

Table 8. A4C view - Video sequences only data

Model	***Frames Size***	***Mean AUC***
ResNet18	112x112	**64.52%**
EchoNet pre-trained	112x112	50.63%
ResNet18	224x224	51.85%
I3D	224x224	52.62%

Table 9. A4C view - Video sequences + Tabular Data

Model	***Frames Size***	***Mean AUC***
ResNet18	112x112	**63.49%**
ResNet18	224x224	58.62%
I3D	224x224	46.03%

In **A4C** view, ResNet18 model at 112x112 achieved the best performance, with mean AUC score of 64.52% in video sequences-only data and 63.49% in combination with tabular data. At 224x224 video sequences-only data, Resnet18 achieved 51.85% mean AUC and 58.62% when tabular data added. The I3D model using 224x224 video sequences-only data achieved lower mean AUC compared to ResNet18, 52.62% and 46.03% when combined with LVEF and LV_GLS values. Lastly, the EchoNet pre-trained model that has been trained on A4C echocardiographic videos as we described in section 2.3.3, fine-tuned using our 112x112 video sequences-only data and achieved a mean AUC score of 50.63%.

4.6 Discussion

In the overall assessment of model performance across all experimental settings, 3D architectures demonstrated promising performance in cardiotoxicity risk prediction in breast cancer patients from echocardiographic video sequences. Among all the configurations, ResNet18 model trained on A4C view using 112x112 data, achieved the highest mean AUC score of 64.52% indicating that in this view low resolution spatiotemporal representations can be highly informative. Similarly, the same architecture when combined with tabular data showed stable performance with mean AUC score of 63.49%, verifying the robustness of this configuration. In contrast, the I3D model, despite its specialized spatiotemporal learning, consistently underperformed ResNet18 in most scenarios, where its highest performance (mean AUC score of 55.12%.) achieved using A2C view in combination with tabular data. EchoNet pre-trained model despite its trained on echocardiographic video data, showed poor performance when fine-tuned in our dataset.

However, the effect of data resolution was not the same across the experiments. For instance, while in A2C view 224x224 provided a small improvement for ResNet18 model in video sequences-only data (55.80% in 112x112 and 56.81% in 224x224), in the A4C view the 112x112 resolution markedly outperformed 224x224 size data with mean AUC score of 64.52%. These findings indicate that the effect of resolution in models' performance varies by echocardiographic view, where the A4C view seems to benefit from lower resolution inputs likely because less detailed frames allow the model to focus better on heart motion instead of being distracted by unnecessary structures.

The integration of tabular produced inconsistent results. Specifically, in some cases such as ResNet18 using 224x224 data in A4C view, the combination of video sequences-only data with tabular data improved the mean AUC score (from 51.85% to 58.62%), while in others it reduced the performance, as noticed for ResNet18 using the same data resolution in A2C view (from 56.81% to 50.34%) or I3D model in A4C view (from 52.62% to 46.03%). These findings suggest that effectiveness of tabular data integration into video sequencies data (multimodal integration) depends on both echocardiographic view and model architecture.

Moreover, we compared our results with deep learning models trained on echocardiographic static images instead of video data. Models achieved an AUC score of 60%

showing that 3D heart video sequences data could enhance AI models performance, especially in predicting cardiotoxicity.

Overall, our results highlight the robust performance of 3D ResNet18 compared to I3D and EchoNet pre-trained models, especially in video sequences-only scenario. In addition, they underscore the importance of experimental settings such as input resolution and multimodal integration, since different echocardiographic views respond differently to model's performance. Lastly, our findings suggest that DL models handling full 3D video sequences can be a useful decision-support tool in cardiotoxicity monitoring by supporting clinical assessments in cardio-oncology practice.

Effect of Spatial Resolution (ResNet18 model only)

3D ResNet18 architecture was trained in both video sequences resolutions, 112x112 and 224x224. Our findings showed that the effect of input size is not the same across echocardiographic views A2C and A4C. In the **A2C** view, the resolution increment from 112x112 to 224x224 showed small performance improvement in video sequences-only scenario (from 55.80% to 56.81%). However, when the tabular data was added, this improvement disappeared compared to first scenario (mean AUC score 56.05% in 112x112 and 50.34% in 224x224). Nevertheless, in the **A4C** view, lower resolution (112x112) led to better performance compared to 224x224 resolution, achieving the highest mean AUC of 64.52%. These findings suggest that in A4C view video sequences, spatial representations are more sufficient and advantageous for cardiac motion patterns capturing relevant to cardiotoxicity. In contrast, higher resolution data may introduce additional noise that hinders model generalization.

ResNet18 and I3D Performance Comparison in 224x224 Data

Using the same spatial resolution, the ResNet18 model outperformed I3D in most scenarios. In the **A2C** view, using video sequences-only data, ResNet18 achieved a mean AUC score of 56.81% while I3D achieved only 46.26%. Nevertheless, in **A4C** view, using the same input, ResNet18 and I3D model achieved 51.85% and 52.62% respectively. In the second scenario, with tabular data integration, the I3D model performance improved in **A2C** view outperforming ResNet18 with mean AUC score of 50.34%, but in **A4C** view I3D's performance decreased to 46.03% compared to ResNet18 that achieved 58.62% in this view. These findings suggest that

ResNet18 model seems to have more stable performance across views, whereas the I3D is more sensitive to echocardiographic view and data input. Despite the fact that I3D is especially designed for spatiotemporal feature learning, in our work did not consistently translate into better classification outcomes. The relatively small dataset that we used, or the clinical variability could limited its performance to capture heart motion patterns effectively, while ResNet18's simpler architecture showed better generalization.

ResNet18 and EchoNet pre-trained Performance Comparison in 112x112 Data

In the **A4C** view using the same spatial resolution (112x112), ResNet18 achieved a mean AUC of 64.52% outperforming EchoNet pre-trained which was fined-tuned on the same data, reached only 50.63%. Despite that EchoNet pre-trained was also trained on A4C echocardiographic videos, was not specifically trained on breast cancer therapy-induced cardiotoxicity data, which could explain its poor ability to generalize in our dataset. Furthermore, we support that differences in imaging protocols, patient population or acquisition settings between EchoNet-Dynamic and CARDIOCARE datasets, further reduced the transferability of its learned features, resulting in the above poor performance compared to ResNet18 model that was trained exclusively from scratch in our data.

Impact of Tabular Data Integration

ResNet18 and I3D models trained in both scenarios, video sequences-only data and the combination of ultrasound video sequences with their corresponding tabular data LVEF and LV_GLS values. The integration of tabular data showed inconsistent results. Specifically, in the **A2C** view, tabular data benefited I3D model, by increasing its mean AUC score from 46.26% to 55.12% but for the ResNet18 model, it reduced its performance from 56.81% to 50.34% at 224x224 and without any significant effect in 112x112 data resolution (from 55.80% to 56.05%). In the **A4C** view, tabular data improved ResNet18's performance from 51.85% to 58.62% at 224x224 but decreased it in 112x112 data from 64.52% to 63.49% and in I3D model from 52.62% to 46.03% using 224x224 data. Based on these results we figure out that the usefulness of tabular features integration depends on both echocardiographic view and model architecture. For ResNet18, the integration of tabular data showed that can improve model's performance in 224x224 spatial resolution, whereas at lower resolution (112x112) video sequences-only features

alone were sufficient. For the I3D, the addition of tabular data showed no performance improvement in A4C while appeared beneficial in A2C view data. These results show that for A2C view tabular data can match well with spatiotemporal features, boosting model performance, but in A4C view tabular data seems to add extra noise and reduce performance.

5. Conclusion

This master's thesis presented two AI applications in medical imaging field, focusing on prostate cancer management and cardiotoxicity risk prediction. Both studies investigated the integration of deep learning methods with medical imaging and clinical data, aiming to address current challenges in precision medicine.

In the first task, we investigated the significance of harmonization and pre-processing techniques in AI models performance enhancement using heterogenous PCa image data, in a binary classification task. Across all experimental settings, 3D model architectures such as ResNet18, demonstrated promising performance in distinguishing between clinically and non-clinically significant PCa. Our findings showed the significant role of harmonization techniques before model training, with histogram and z-score normalization offering model generalizability enhancement and inter-site variability reduction. On the other hand, techniques such as the combination of N4 bias field correction with z-score normalization showed poor performance and inconsistent results, highlighting the need for careful pre-preprocessing pipelines selection. In addition, our outcomes demonstrated the impact of spatial resolution, where larger volumes indeed include more clinical anatomical features but also introduce computational complexity. Lastly, despite the imbalanced data, harmonization techniques improved overall models' performance, showing their value in multi-centric and multi-vendor PCa data.

In the second task, we assessed the performance of 3D AI models for cardiotoxicity risk prediction using full 3D video sequences with and without the integration of tabular data, acquired from breast cancer patients. ResNet18 achieved the best performance, particularly in the A4C view using the lower resolution input, demonstrating that even spatiotemporal representations in low resolution can effectively capture cardiac motion patterns related to cardiotoxicity. In contrast, the I3D model despite its specialized design for spatiotemporal features learning, showed poor performance in most settings likely due to dataset limitations and clinical variability. Moreover, multimodal integration with clinical tabular data led to mixed results, benefitting the model's performance in certain echocardiographic views but harming in others. These findings showed that both view selection and input resolution have an important role in experimental design, as they influenced model's performance. Also the fine-tuning of EchoNet pre-trained model in our data,

showed poor prediction performance, despite the fact that was originally trained on A4C echocardiographic data.

Considering both tasks, this study shows the potential of deep learning as a decision-support tool in oncology and cardiology fields, but we recognize that several challenges still exist. Nevertheless, we anticipate that the findings arising from this master's thesis will further contribute to prostate cancer management and cardiotoxicity risk prediction research.

5.1 Study Limitations

The findings of this master's thesis demonstrate the potential of AI models as a decision-support tool in prostate management and cardiotoxicity risk prediction but simultaneously unfold some limitations that must be considered in future works. The datasets that we used in both tasks were provided by the two EU projects ProCAncer-i and CARDIOCARE, offering unique imaging data, such as heterogeneous PCa MRI images and heart ultrasound video sequences acquired along with their clinical data from breast cancer patients respectively. However, for the cardiotoxicity risk prediction task the number of samples was relatively small especially after the pre-processing phase (N=256 for A4C and N=201 for A2C). Similarly, for the PCa management task, the provided data were highly imbalanced (N=1102 for non-clinically significant and N=2833 for clinically significant). Taken together, these limitations introduce concerns about overall model's performance and generalizability in real clinical practice. Moreover, due to the clinical assessment, regarding the data collection in cardiotoxicity risk prediction task, is still ongoing, the current outcomes could be change in the future as new data are added or as the status of a breast cancer patient could be also change, e.g. a classified patient as non-cardiotoxic may develop cardiotoxicity in a few months as the cancer treatment continues.

5.2 Future Work

Despite the promising results of this study, we recognize that further investigations are needed in future research. Regarding the prostate cancer management task, future directions could be focus on exploring other harmonization techniques and deep learning architectures to reduce further the inter-site variability between different clinical centers and machine vendors. It is also crucial for the class imbalance to be addressed, by collecting more balanced cohorts or through

data generation methods such as AI generated synthetic images (using GANs) to increase the balance of the dataset and models' generalizability. In addition, the integration of clinical data such as patient age, blood biomarkers including PSA value, could increase model's predictive capability and clinical support. Concerning the cardiotoxicity risk prediction task, future approaches could focus on different deep learning models' implementations to assess their effectiveness in capturing heart motion patterns from unique video sequences data acquired from breast cancer patients undergoing cancer treatment. The addition of blood biomarkers alongside LVEF and LV_GLS values, treatment-related data or patient demographic information, could also increase models' overall performance and generalizability. Moreover, the inclusion of echocardiographic views beyond A2C and A4C could also lead to better understanding of cardiac function.

References

[1] A. Barragán-Montero *et al.*, "Artificial intelligence and machine learning for medical imaging: A technology review," Mar. 01, 2021, *Associazione Italiana di Fisica Medica*. doi: 10.1016/j.ejmp.2021.04.016.

[2] J. Olveres *et al.*, "What is new in computer vision and artificial intelligence in medical image analysis applications," Aug. 01, 2021, *AME Publishing Company*. doi: 10.21037/qims-20-1151.

[3] I. Castiglioni *et al.*, "AI applications to medical images: From machine learning to deep learning," Mar. 01, 2021, *Associazione Italiana di Fisica Medica*. doi: 10.1016/j.ejmp.2021.02.006.

[4] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," Dec. 2015, [Online]. Available: http://arxiv.org/abs/1511.08458

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition." [Online]. Available: http://image-net.org/challenges/LSVRC/2015/

[6] D. Sarwinda, R. H. Paradisa, A. Bustamam, and P. Anggia, "Deep Learning in Image Classification using Residual Network (ResNet) Variants for Detection of Colorectal Cancer," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 423–431. doi: 10.1016/j.procs.2021.01.025.

[7] V.-T. Hoang and H. Jo, "Practical Analysis on Architecture of EfficientNet."

[8] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 2015, [Online]. Available: http://arxiv.org/abs/1409.1556

[9] M. M. Rahman, M. S. Islam, R. Sassi, and M. Aktaruzzaman, "Convolutional neural networks performance comparison for handwritten Bengali numerals recognition," *SN Appl Sci*, vol. 1, no. 12, Dec. 2019, doi: 10.1007/s42452-019-1682-y.

[10] S. Majumdar, P. Pramanik, and R. Sarkar, "Gamma function based ensemble of CNN models for breast cancer detection in histopathology images," *Expert Syst Appl*, vol. 213, Mar. 2023, doi: 10.1016/j.eswa.2022.119022.

[11] V. Iglovikov and A. Shvets, "TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation," Jan. 2018, [Online]. Available: http://arxiv.org/abs/1801.05746

[12] J. Carreira and A. Zisserman, “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset,” Feb. 2018, [Online]. Available: http://arxiv.org/abs/1705.07750

[13] A. Vaswani *et al.*, “Attention Is All You Need.”

[14] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” Jun. 2021, [Online]. Available: http://arxiv.org/abs/2010.11929

[15] K. Han *et al.*, “A Survey on Vision Transformer,” *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 1, pp. 87–110, Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.

[16] L. M. Prevedello *et al.*, “Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions,” *Radiol Artif Intell*, vol. 1, no. 1, Jan. 2019, doi: 10.1148/ryai.2019180031.

[17] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” *J Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0197-0.

[18] H. Talebi, P. Milanfar, and G. Research, “Learning to Resize Images for Computer Vision Tasks.”

[19] F. Schwarzhans *et al.*, “Image normalization techniques and their effect on the robustness and predictive power of breast MRI radiomics,” *Eur J Radiol*, vol. 187, Jun. 2025, doi: 10.1016/j.ejrad.2025.112086.

[20] M. Butwall, “Data Normalization and Standardization: Impacting Classification Model Accuracy,” 2021.

[21] A. Demircioğlu, “The effect of feature normalization methods in radiomics,” *Insights Imaging*, vol. 15, no. 1, Dec. 2024, doi: 10.1186/s13244-023-01575-7.

[22] M. Butwall, “Data Normalization and Standardization: Impacting Classification Model Accuracy,” 2021.

[23] A. Demircioğlu, “The effect of feature normalization methods in radiomics,” *Insights Imaging*, vol. 15, no. 1, Dec. 2024, doi: 10.1186/s13244-023-01575-7.

[24] E. Stamoulou *et al.*, “Harmonization Strategies in Multicenter MRI-Based Radiomics,” Nov. 01, 2022, *MDPI*. doi: 10.3390/jimaging8110303.

[25] L. G. Nyú and J. K. Udupa, “On Standardizing the MR Image Intensity Scale,” 1999.

[26] S. Seoni *et al.*, “All you need is data preparation: A systematic review of image harmonization techniques in Multi-center/device studies for medical support systems,” Jun. 01, 2024, *Elsevier Ireland Ltd*. doi: 10.1016/j.cmpb.2024.108200.

[27] B. J. Erickson and F. Kitamura, “Magician’s corner: 9. performance metrics for machine learning models,” May 01, 2021, *Radiological Society of North America Inc.* doi: 10.1148/ryai.2021200126.

[28] M. Bramer, *Principles of Data Mining*. in Undergraduate Topics in Computer Science. London: Springer London, 2020. doi: 10.1007/978-1-4471-7493-6.

[29] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, “The balanced accuracy and its posterior distribution,” in *Proceedings - International Conference on Pattern Recognition*, 2010, pp. 3121–3124. doi: 10.1109/ICPR.2010.764.

[30] P. Rawla, “Epidemiology of Prostate Cancer,” *World J Oncol*, vol. 10, no. 2, pp. 63–89, 2019, doi: 10.14740/wjon1191.

[31] M. Sekhoacha, K. Riet, P. Motloung, L. Gumenku, A. Adegoke, and S. Mashele, “Prostate Cancer Review: Genetics, Diagnosis, Treatment Options, and Alternative Approaches,” Sep. 01, 2022, *MDPI*. doi: 10.3390/molecules27175730.

[32] S. W. D. Merriel *et al.*, “Systematic review and meta-analysis of the diagnostic accuracy of prostate-specific antigen (PSA) for the detection of prostate cancer in symptomatic patients,” *BMC Med*, vol. 20, no. 1, Dec. 2022, doi: 10.1186/s12916-021-02230-y.

[33] P. Porzycki and E. Ciszkowicz, “Modern biomarkers in prostate cancer diagnosis,” 2020, *Polish Urological Association*. doi: 10.5173/ceju.2020.0067R.

[34] G. Pirtskhalaishvili, R. L. Hrebinko, and J. B. Nelson, “The Treatment of Prostate Cancer. ,” *Cancer Pract*, vol. 9, no. 6, pp. 295–306, Nov. 2001, doi: 10.1111/j.1523-5394.2001.96009.pp.x.

[35] A. Berti *et al.*, “Exploring the potentials and challenges of Artificial Intelligence in supporting clinical diagnostics and remote assistance for the health and well-being of individuals,” 2022. [Online]. Available: http://ceur-ws.org

[36] G. Maimone, G. Nicoletti, S. Mazzetti, D. Regge, and V. Giannini, “Comparison of Machine and Deep Learning models for automatic segmentation of prostate cancers on multiparametric MRI,” in *2022 IEEE International Symposium on Medical Measurements and Applications, MeMeA 2022 - Conference Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/MeMeA54994.2022.9856530.

[37] Ö. Önder *et al.*, “Long-term follow-up results of multiparametric prostate MRI and the prognostic value of PI-RADS: a single-center retrospective cohort study,” *Diagnostic and*

Interventional Radiology, vol. 30, no. 3, pp. 139–151, May 2024, doi: 10.4274/dir.2023.232414.

[38] S. Masoudi *et al.*, "Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research," *Journal of Medical Imaging*, vol. 8, no. 01, Jan. 2021, doi: 10.1117/1.jmi.8.1.010901.

[39] J. Wu, "The Application of Artificial Intelligence Image Analysis Techniques in Assisting with the Classification and Preprocessing of MRI Prostate Cancer Diagnostic Data," in *2024 5th International Conference on Electronic Communication and Artificial Intelligence, ICECAI 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 401–405. doi: 10.1109/ICECAI62591.2024.10675240.

[40] P. M. Ebin, U. Ananthanagu, G. C. Mara, B. Indu, R. Thomas, and N. M. Mathkunti, "A Breakthrough Approach for Prostate Cancer Identification Utilizing VGG-16 CNN Model with Migration Learning," in *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation, IATMSI 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/IATMSI60426.2024.10503011.

[41] D. Zaridis, E. Mylona, N. Tachos, K. Marias, M. Tsiknakis, and D. I. Fotiadis, "A Deep Learning-based cropping technique to improve segmentation of prostate's peripheral zone," in *BIBE 2021 - 21st IEEE International Conference on BioInformatics and BioEngineering, Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/BIBE52308.2021.9635576.

[42] D. I. Zaridis *et al.*, "Region-adaptive magnetic resonance image enhancement for improving CNN-based segmentation of the prostate and prostatic zones," *Sci Rep*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-27671-8.

[43] A. Saha, M. Hosseinzadeh, and H. Huisman, "End-to-end prostate cancer detection in bpMRI via 3D CNNs: Effects of attention mechanisms, clinical priori and decoupled false positive reduction," *Med Image Anal*, vol. 73, Oct. 2021, doi: 10.1016/j.media.2021.102155.

[44] A. M. Ali and A. A. Mohammed, "Improving classification accuracy for prostate cancer using noise removal filter and deep learning technique," *Multimed Tools Appl*, vol. 81, no. 6, pp. 8653–8669, Mar. 2022, doi: 10.1007/s11042-022-12102-z.

[45] S. Liu, H. Zheng, Y. Feng, and W. Li, “Prostate cancer diagnosis using deep learning with 3D multiparametric MRI,” in *Medical Imaging 2017: Computer-Aided Diagnosis*, SPIE, Mar. 2017, p. 1013428. doi: 10.1117/12.2277121.

[46] H. Sherafatmandjoo, A. A. Safaei, F. Ghaderi, and F. Allameh, “Prostate cancer diagnosis based on multi-parametric MRI, clinical and pathological factors using deep learning,” *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-65354-0.

[47] S. Yoo, I. Gujrathi, M. A. Haider, and F. Khalvati, “Prostate Cancer Detection using Deep Convolutional Neural Networks,” *Sci Rep*, vol. 9, no. 1, Dec. 2019, doi: 10.1038/s41598-019-55972-4.

[48] N. Aldoj, S. Lukas, M. Dewey, and T. Penzkofer, “Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network,” *Eur Radiol*, vol. 30, no. 2, pp. 1243–1253, Feb. 2020, doi: 10.1007/s00330-019-06417-z.

[49] C. De Vente, P. Vos, M. Hosseinzadeh, J. Pluim, and M. Veta, “Deep Learning Regression for Prostate Cancer Detection and Grading in Bi-Parametric MRI,” *IEEE Trans Biomed Eng*, vol. 68, no. 2, pp. 374–383, Feb. 2021, doi: 10.1109/TBME.2020.2993528.

[50] M. R. Hassan *et al.*, “Prostate cancer classification from ultrasound and MRI images using deep learning based Explainable Artificial Intelligence,” *Future Generation Computer Systems*, vol. 127, pp. 462–472, Feb. 2022, doi: 10.1016/j.future.2021.09.030.

[51] E. Pachetti, S. Colantonio, and M. A. Pascali, “On the Effectiveness of 3D Vision Transformers for the Prediction of Prostate Cancer Aggressiveness,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Science and Business Media Deutschland GmbH, 2022, pp. 317–328. doi: 10.1007/978-3-031-13324-4_27.

[52] S. G. Armato *et al.*, “PROSTATEx Challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images,” *Journal of Medical Imaging*, vol. 5, no. 04, p. 1, Nov. 2018, doi: 10.1117/1.jmi.5.4.044501.

[53] J. Caron and A. Nohria, “Cardiac Toxicity from Breast Cancer Treatment: Can We Avoid This?,” Aug. 01, 2018, *Current Medicine Group LLC 1*. doi: 10.1007/s11912-018-0710-1.

[54] A. Albini, G. Pennesi, F. Donatelli, R. Cammarota, S. De Flora, and D. M. Noonan, "Cardiotoxicity of anticancer drugs: The need for cardio-oncology and cardio-oncological prevention," Jan. 2010. doi: 10.1093/jnci/djp440.

[55] C. Papageorgiou, A. Andrikopoulou, M. A. Dimopoulos, and F. Zagouri, "Cardiovascular toxicity of breast cancer treatment: an update," Jul. 01, 2021, *Springer Science and Business Media Deutschland GmbH*. doi: 10.1007/s00280-021-04254-w.

[56] J. D. Floyd, D. T. Nguyen, R. L. Lobins, Q. Bashir, D. C. Doll, and M. C. Perry, "Cardiotoxicity of cancer therapy," 2005. doi: 10.1200/JCO.2005.08.789.

[57] M. S. Ewer and S. M. Ewer, "Cardiotoxicity of anticancer treatments," Sep. 21, 2015, *Nature Publishing Group*. doi: 10.1038/nrcardio.2015.65.

[58] H. Kim *et al.*, "Diagnosis, treatment, and prevention of cardiovascular toxicity related to anti-cancer treatment in clinical practice: an opinion paper from the working group on cardio-oncology of the Korean society of echocardiography," 2018, *Korean Society of Echocardiography*. doi: 10.4250/jcu.2018.26.1.1.

[59] P. P. Zheng, J. Li, and J. M. Kros, "Breakthroughs in modern cancer therapy and elusive cardiotoxicity: Critical research-practice gaps, challenges, and insights," Jan. 01, 2018, *John Wiley and Sons Inc.* doi: 10.1002/med.21463.

[60] A. Seraphim *et al.*, "Advanced Imaging Modalities to Monitor for Cardiotoxicity," *Curr Treat Options Oncol*, vol. 20, no. 9, Sep. 2019, doi: 10.1007/s11864-019-0672-z.

[61] W. Cao *et al.*, "Clinical Challenges and AI Opportunities in Decision-Making for Cancer Treatment-Induced Cardiotoxicity," ACM, 2024. doi: XXXXXXX.XXXXXXX.

[62] beIbeI JIang, nIng guo, Yi. ge, lu Zhang, matthIJs ouDkerk, and X. XIe, "Imaging Patients with stable chest pain special feature: review article Development and application of artificial intelligence in cardiac imaging," 2020. [Online]. Available: https://academic.oup.com/bjr/article/93/1113/20190812/7451994

[63] K. Lekadir, T. Leiner, A. A. Young, and S. E. Petersen, "Editorial: Current and Future Role of Artificial Intelligence in Cardiac Imaging," Aug. 07, 2020, *Frontiers Media S.A.* doi: 10.3389/fcvm.2020.00137.

[64] B. Liu *et al.*, "A deep learning framework assisted echocardiography with diagnosis, lesion localization, phenogrouping heterogeneous disease, and anomaly detection," *Sci Rep*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-022-27211-w.

[65] W. T. Chang *et al.*, “An artificial intelligence approach for predicting cardiotoxicity in breast cancer patients receiving anthracycline,” *Arch Toxicol*, vol. 96, no. 10, pp. 2731–2737, Oct. 2022, doi: 10.1007/s00204-022-03341-y.

[66] D. J. Choi, J. J. Park, T. Ali, and S. Lee, “Artificial intelligence for the diagnosis of heart failure,” *NPJ Digit Med*, vol. 3, no. 1, Dec. 2020, doi: 10.1038/s41746-020-0261-3.

[67] A. Madani, R. Arnaout, M. Mofrad, and R. Arnaout, “Fast and accurate view classification of echocardiograms using deep learning,” *NPJ Digit Med*, vol. 1, no. 1, Dec. 2018, doi: 10.1038/s41746-017-0013-1.

[68] S. Magesh and P. S. Rajakumar, “FETAL HEART DISEASE DETECTION VIA DEEP REG NETWORK BASED ON ULTRASOUND IMAGES.”

[69] J. Zhang *et al.*, “Fully automated echocardiogram interpretation in clinical practice: Feasibility and diagnostic accuracy,” *Circulation*, vol. 138, no. 16, pp. 1623–1635, 2018, doi: 10.1161/CIRCULATIONAHA.118.034338.

[70] J. P. Howard *et al.*, “Improving ultrasound video classification: An evaluation of novel deep learning methods in echocardiography,” *J Med Artif Intell*, vol. 3, no. March, Mar. 2020, doi: 10.21037/jmai.2019.10.03.

[71] A. Talebi *et al.*, “Machine learning based radiomics model to predict radiotherapy induced cardiotoxicity in breast cancer,” *J Appl Clin Med Phys*, vol. 26, no. 4, Apr. 2025, doi: 10.1002/acm2.14614.

[72] K. M. Tsiouris *et al.*, “Risk Stratification for Cardiotoxicity in Breast Cancer Patients: Predicting Early Decline of LVEF After Treatment,” in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/EMBC40787.2023.10340316.

[73] D. Ouyang *et al.*, “Video-based AI for beat-to-beat assessment of cardiac function,” *Nature*, vol. 580, no. 7802, pp. 252–256, Apr. 2020, doi: 10.1038/s41586-020-2145-8.

[74] N. M. Rodrigues *et al.*, “Analysis of domain shift in whole prostate gland, zonal and lesions segmentation and detection, using multicentric retrospective data,” *Comput Biol Med*, vol. 171, Mar. 2024, doi: 10.1016/j.compbiomed.2024.108216.

[75] F. Pérez-García, R. Sparks, and S. Ourselin, “TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning,” Aug. 2021, doi: 10.1016/j.cmpb.2021.106236.

[76] C. R. Harris *et al.*, “Array programming with NumPy,” Sep. 17, 2020, *Nature Research*. doi: 10.1038/s41586-020-2649-2.

[77] N. J. Tustison *et al.*, “The ANTsX ecosystem for quantitative biological and medical imaging,” *Sci Rep*, vol. 11, no. 1, p. 9068, Apr. 2021, doi: 10.1038/s41598-021-87564-6.

[78] M. J. Cardoso *et al.*, “MONAI: An open-source framework for deep learning in healthcare,” Nov. 2022, [Online]. Available: http://arxiv.org/abs/2211.02701

[79] T. Kluyver *et al.*, “Jupyter Notebooks—a publishing format for reproducible computational workflows,” in *Positioning and Power in Academic Publishing: Players, Agents and Agendas - Proceedings of the 20th International Conference on Electronic Publishing, ELPUB 2016*, IOS Press BV, 2016, pp. 87–90. doi: 10.3233/978-1-61499-649-1-87.

[80] F. Pedregosa FABIANPEDREGOSA *et al.*, “Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot,” 2011. [Online]. Available: http://scikit-learn.sourceforge.net.