Instituto Politécnico de Lisboa

Instituto Superior de Engenharia de Lisboa

Escola Superior de Tecnologia da Saúde de Lisboa

# Convolutional Neural Networks for Myocardial Perfusion SPECT Imaging Classification: A Full and Low-Dose Study

Gonçalo Cravo de Jesus Gonçalves

Thesis to obtain the Master of Science Degree in Biomedical Engineering

**Supervisors**

Sérgio Rafael Reis Figueiredo (ESTeSL)

Pedro Miguel Torres Mendes Jorge (ISEL)

December 2023

Instituto Politécnico de Lisboa

Instituto Superior de Engenharia de Lisboa

Escola Superior de Tecnologia da Saúde de Lisboa

# Convolutional Neural Networks for Myocardial Perfusion SPECT Imaging Classification: A Full and Low-Dose Study

Gonçalo Cravo de Jesus Gonçalves

Thesis to obtain the Master of Science Degree in Biomedical Engineering

**Supervisors**

Sérgio Rafael Reis Figueiredo (ESTeSL)

Pedro Miguel Torres Mendes Jorge (ISEL)

**Examination Committee**

Chairperson: João Pedro Barrigana Ramos da Costa (ISEL)

Member of the Committee: Rita Homem de Gouveia Costanzo Nunes (IST-ULisboa)

Supervisor: Sérgio Rafael Reis Figueiredo (ESTeSL)

December 2023

# Acknowledgements

Terminada esta etapa gostaria de agradecer ao Professor Sérgio Figueiredo por ter aceitado ser meu orientador e por todos os conhecimentos passados nos últimos seis anos no decorrer da minha Licenciatura e Mestrado. Obrigado por todos os conselhos, apoio e paciência.

Ao Doutor Pedro Jorge, agradeço também todos os conhecimentos passados no decorrer do Mestrado e na realização desta tese. Obrigado pelo seu constante apoio e diferentes pontos de vista, que me ensinaram a ser mais crítico no meu trabalho e nos problemas que enfratava.

Queria agradecer também à NuclearMed, Hospital Particular de Almada (HPA), por ter fornecido os estudos cardíacos para a realização deste trabalho. Adicionalmente, um especial obrigado ao Professor e Técnico de Medicina Nuclear Edgar Pereira por toda a disponibilidade na reconstrução dos estudos que também fez com que este trabalho fosse possível.

Aos meus colegas e amigos de Mestrado, em particular à Ana Beatriz Mateus, Diogo Barradas, Filipa Jacques e Ricardo Monteiro, por todo o apoio durante estes dois anos e por todos os momentos felizes que passámos.

Aos meus amigos mais chegados, em particular à Catarina R., Cláudia S., Daniela D., Daniela S., Daniela T., Inês S., Maria F., Miguel F. e Sofia C., por todo o apoio e amizade incondicional dos últimos seis anos. Obrigado por estarem sempre por perto.

Por último, à minha família pelo carinho e por sempre terem acreditado em mim. Obrigado à minha mãe, ao meu pai, ao meu irmão Guilherme e ao Galileu. A vós dedico este trabalho.

# Abstract

Myocardial perfusion imaging (MPI) by single-photon emission computed tomography (SPECT) plays a crucial role in the diagnosis of coronary artery disease. Moreover, the quantification of these images typically involves the extraction of quantitative parameters obtained from the rest stress perfusion. However, the acquisition systems have some limitations such as spatial blurring and low-count data, which may introduce bias in the classification. Additionally, these processes are time-consuming and user-dependent, leading to significant intra and inter-operator variability. Furthermore, over the years there has been a constant effort to reduce the dose of MPI.

In this sense, both the variability classification issues and the dose reduction concerns can impact the true assessment of SPECT-MPI. In recent years, with the rise of artificial intelligence algorithms, several studies have proposed automatic Deep Learning techniques for the classification of MPI, moreover regarding low-count data.

In this project, we ran 5 Convolutional Neural Network models with pre-trained weights: one trained on stress real full-time data (100%, as 100R), three individual models with synthetic 75%, 50% and 25% count settings and another one with all datasets combined (ALL). Thus, we compared their performance when tested on full-time and low-time studies and assessed the application of synthetic subsampled data from the Poisson Resampling technique in SPECT-MPI classification tasks.

As a conclusion, both 100R and ALL models achieved good and similar results when tested with real full-time (100R model achieved an accuracy of 0.70 and ALL model achieved an accuracy of 0.65) and real low-time at 75% (both models achieved and accuracy of 0.71). Bellow this percentage, the models' accuracy began to drop, possibly due to the limited information these images contain. Thus, subsampled data from a Poisson resampling method may be a possible solution to conduct further studies regarding the classification of low-time SPECT-MPI.

*Keywords: Deep Learning, Convolutional Neural Networks, Image Classification SPECT, Myocardial Perfusion Imaging*

# Resumo

As doenças cardiovasculares são a principal causa de morbilidade e mortalidade a nível mundial. A doença arterial coronária, uma das formas mais prevalentes das doenças cardiovasculares, caracteriza-se pela deposição de placas de ateroma nas artérias coronárias, levando consequentemente a diminuições da passagem do sangue pelas artérias, fenómeno este denominado por isquémia. As isquémias podem ser reversíveis, quando existe uma diminuição da passagem do sangue temporária, causada por exercício físico ou testes de esforço farmacológico ou irreversíveis quando o estreitamento e oclusão do vaso não é revertido em situações de repouso, levando à necrose dos tecidos. Neste sentido, a cintigrafia de perfusão do miocárdio (CPM) por tomografia computorizada por emissão de fotão único (SPECT) é um exame não-invasivo muito utilizado e designado para detetar alterações de perfusão no músculo cardíaco.

Na prática, a SPECT utiliza traçadores radioativos que depois de administrados no doente permitem uma visualização do mesmo pelo corpo ou órgão de interesse. Na CPM, após administrado o radiofármaco, este é fixado parcialmente no miocárdio proporcionalmente ao fluxo sanguíneo do músculo. Na prática, em áreas indicativas de baixa perfusão, pode ser indicativo de isquémia reversível ou irreversível (enfarte). Se a região lesada se apresentar menos perfundida num estudo em esforço que num estudo em repouso, estaremos perante uma isquémia reversível. Se a região não apresentar reversibilidade de perfusão no estudo em repouso, estamos perante de uma isquémia irreversível e, portanto, um enfarte. Neste sentido, na CPM, é a comparação entre o estudo em esforço e em repouso que indica a presença ou ausência de estenoses.

A quantificação dos estudos de CPM-SPECT envolve tipicamente a extração de parâmetros quantitativos obtidos dos estudos de esforço e repouso. Para isso, são utilizados softwares comercialmente disponíveis como o QPS$^{TM}$/QGS$^{TM}$ e 4D-MSPECT. No entanto, os sistemas de aquisição sofrem de algumas limitações como a distorção espacial, atenuação dispersa e baixas

contagens, o que pode contribuir para elevados níveis de ruído nas imagens. Assim, este aspeto pode levar ao viés na quantificação e consequente classificação, para além de serem processos demorados e muito dependentes do utilizador, levando a variabilidades intra e inter-operador significantes.

Além disso, a CPM-SPECT sendo um exame que comporta radiação ionizante para o doente, é o que mais contribui em Portugal para a dose coletiva da população desde 2010. Para ultrapassar tais limitações, novos equipamentos especializados em imagem cardíaca nuclear têm sido desenvolvidos como os de Cádmio-zinco-telureto (CZT), permitindo uma melhor qualidade de imagem e redução de dose. Outra abordagem foi implementar algoritmos iterativos na reconstrução de imagem, melhorando também a relação sinal-ruído das mesmas. No entanto, ambos os métodos estão associados a custos elevados, sendo difíceis de implementar em muitos serviços.

Neste sentido, tanto as questões na variabilidade de classificação como as preocupações de redução da dose podem impactar a verdadeira avaliação de estudos CPM-SPECT. Recentemente, com a ascensão do *big-data* e novos algoritmos de aprendizagem profunda como as redes neuronais, estes começaram a ser utilizados também no contexto da imagiologia médica, particularmente também na CPM. Para contexto, muitas arquiteturas de redes neuronais têm sido utilizadas nas mais variadas tarefas da imagiologia, sendo as redes neuronais convolucionais a que mais se destacou. Uma razão para tal, é a facilidade da sua utilização através de *transfer learning*, ou seja, a utilização de redes neuronais convolucionais pré-treinadas numa base de dados de grandes dimensões, como por exemplo o ImageNet, em que se congela os pesos das camadas anteriores ou parte deles, de forma a usar o que a rede aprendeu inicialmente num outro problema de classificação.

Em adendo, com o desenvolvimento de arquiteturas cada vez mais complexas, a interpretabilidade destes modelos tornou-se mais difícil. Para isso, vários métodos foram desenvolvidos como forma de melhor perceber a tomada de decisão dos algoritmos nas mais variadas tarefas, dentro dos quais destacamos o Grad-CAM.

Neste sentido, e para contexto, as redes neuronais convolucionais têm sido utilizadas na cardiologia nuclear nas mais variadas tarefas como na otimização da qualidade de imagens, na reconstrução de dados e na classificação de estudos para apoio ao diagnóstico. Adicionalmente, vários investigadores têm estudado particularmente a classificação em imagens de baixa dose depois de sobreamostragem para dose convencional. No entanto, a literatura ainda carece de

estudos relativos à classificação de imagens de dose convencional e reduzida com posterior comparação entre ambas.

Neste projeto temos como objetivo desenvolver um modelo de classificação de imagens reais *full-time* e *low-time* de perfusão do miocárdio, avaliando a aplicação do uso de imagens sintéticas geradas a partir do método de reamostragem de Poisson em tarefas de classificação.

**Métodos**

Este trabalho inclui um *cohort* de 129 doentes admitidos para CPM. Desses, adicionalmente às aquisições de tempo convencional, 21 estudos foram re-adquiridos com proporções de 75%, 50% e 25% do tempo convencional de aquisição. Para aumentar o número da amostra de imagens e como tarefas de classificação com redes neuronais geralmente precisam de grandes quantidades de dados, aplicámos a reamostragem de Poisson em toda a amostra, sintetizando imagens de tempo reduzido a 75%, 50% e 25%. Neste sentido, foram treinados cinco modelos da ResNet-18 com pesos pré-treinados em ImageNet: um com dados reais de esforço pós-reconstruídos com tempo convencional (100R), três individuais com dados sintéticos a 75%, 50% e 25% do tempo convencional (75S, 50S e 25S) e um último com estes quatro conjuntos de dados combinados (ALL). Como objetivo, cada modelo aprende a classificar imagens reais a *full-time* e reais e sintéticas a *low-time* (75%, 50% e 25%). Para avaliar o desempenho dos modelos, usámos as métricas taxa de acerto, sensibilidade, especificidade, *f1-score* e a área abaixo da curva (AUC) derivada da *receiver operator characteristics curve* (ROC). Para efeitos de interpretabilidade usámos o método Grad-CAM.

**Resultados**

Relativamente ao treino 100R, este atingiu uma taxa de acerto de 0.70 e 0.71 e AUC de 0.73 e 0.61 quando testado com os dados de *full-time* e reais a 75%, respetivamente. No conjunto de teste a 75% sintético, o desempenho baixou ligeiramente atingindo uma taxa de acerto de 0.57, devido à incapacidade de identificar os casos "normais" (especificidade = 0.50). Em percentagens de tempo inferior, os resultados na classificação diminuíram gradativamente. Quanto aos treinos individuais com dados sintéticos, 75S, 50S e 25S, percebemos que estes modelos conseguem classificar dados quando testados com as mesmas percentagens. No treino 25S e 50S obteu-se assim uma taxa de acerto de 0.65 e no treino com 75S obteu-se uma taxa de acerto de 0.62. Finalmente, no treino ALL, os resultados foram bastante similares ao treino

100R, atingindo uma taxa de acerto de 0.71 no grupo de teste a 100%, comparativamente aos 0.70 do treino 100R no mesmo grupo de teste. Relativamente aos grupos de teste real a 75%, atingiu uma taxa de acerto de 0.71 e 0.57 de *f1-score* em ambos os modelos de treino. Nas percentagens mais baixas a 50% e a 25%, os resultados foram de encontro aos obtidos no modelo 100R, em que o desempenho do classificador diminuiu. A partir da análise dos resultados com Grad-CAM, concluímos que apesar dos modelos por vezes classificarem corretamente as imagens tendo em conta cortes mediais relevantes, por outras o modelo leva demasiado em consideração zonas com limitada informação, como por exemplo, cortes cardíacos da base ou ápice. Cremos que isso poderá ser explicado pela forma como organizámos os dados na etapa de preparação das imagens de entrada para os modelos de aprendizagem automática.

**Conclusões**

Como conclusão, o modelo 100R permite a classificar satisfatoriamente imagens reais e sintéticas do estudo de esforço de CPM, particularmente quando testado com dados adquiridos com tempo convencional e com redução a 75%. No entanto, em percentagens de tempo inferiores, o desempenho dos classificadores diminui, eventualmente relacionado com degradação da qualidade da imagem, associada à diminuição do total de contagens não compensada. A introdução de dados com redução de tempo para treino do modelo ALL, parece não incrementar ganho individual por cada grupo, revelando resultados similares ao modelo 100R. A técnica de reamostragem de Poisson aplicada em CPM-SPECT permite obter imagens sintéticas viáveis, contribuindo para eventual minimização da utilização de dados reais, podendo ser aplicada com sucesso em tarefas classificação de estudos CPM-SPECT.

Neste sentido, são necessários trabalhos futuros, particularmente, utilizando uma amostra de maiores dimensões, e cuja divisão de dados para treino e teste permita obter modelos mais robustos e generalizáveis. Adicionalmente, poderá ser adequado utilizar outras arquitecturas de redes neuronais, com caraterísticas otimizadas para o problema da classificação das imagens CPM-SPECT.

*Palavras-chave: Aprendizagem Profunda, Redes Neuronais Convolucionais, Classificação de imagens SPECT, Cintigrafia de Perfusão do Miocárdio*

# Contents

# List of Figures

# List of Tables

# Acronyms

**1D** unidimensional

**2D** bidimensional

**3D** tridimensional

**$^{99}$Mo** Molybdenum 99

**$^{201}$TI** Tallium 201

**$^{99m}$Tc-sestamibi** Technetium 99m-metastable - sestamibi

**$^{99m}$Tc-tetrosfosmin** Technetium 99m-metastable - tetrofosmin

**Adam** Adaptive Moment Estimation

**AI** Artificial Intelligence

**ALARA** As Low as Reasonably Achievable

**ANN** Artificial Neural Networks

**ART** Algebraic Reconstruction Technique

**ASNC** American Association of Nuclear Cardiology

**CAD** Coronary Artery Disease

**CFR** Coronary flow reserve

**CNN** Convolutional Neural Network

**CT** Computed Tomography

**CVD** Cardiovascular disease

**CZT** Cadmium-zinc-telluride

**DICOM** Digital Imaging and Communications in Medicine

**DL** Deep Learning

**DNN** Deep Neural Network

**EDV** end-diastolic volume

**ESV** end-systolic volume

**FBP** Filtered Back Projection

**FBP-W** Filtered Back Projection with Wiener filter

**FDA** Food and Drug Administration

**FN** False Negatives

**FoV** Field of View

**FP** False Positives

**GAN** Generative Adversarial Networks

**GAP** Global Average Polling

**GPU** Graphic Processing Units

**Grad-CAM** Gradient-Weighted Class Activation Mapping

**HIPAA** Health Insurance Portability and Accountability Act

**HLA** Horizontal long-axis

**IAEA** International Atomic Energy Agency

**ILSVRC** ImageNet Large Scale Visual Recognition Challenge

**LEHR** Low energy high resolution

**LOR** Line of Response

**LPO** Left Posterior oblique

**LV** Left ventricle

**LVEF** Left ventricle ejection fraction

**Max** maximum

**MBF** Myocardial Blood Flow

**min** minimum

**ML** Machine Learning

**MLEM** Maximum-Likelihood Expectation Maximization

**MPI** Myocardial Perfusion Imaging

**MRI** Magnetic Imaging Resonance

**NAI(TI)** Tallium-activated sodium iodide

**NN** Neural Networks

**OSEM** Ordered-subsets Expectation Maximization

**RAO** Right anterior oblique

**rMBF** regional Myocardial Blood Flow

**ROC-AUC** Receives operating curve - Area Under the curve

**SA** Short-axis

**SD** standard deviation

**SDS** Summed Difference Score

**SGD** Stochastic Gradient Descents

**SIRT** Simultaneous Reconstruction Technique

**SNR** Signal-to-noise ratio

**SPECT** Single-photon emission computed tomography

**SRS** Summed Rest Score

**SSS** Summed Stress Score

**TN** True Negatives

**TP** True Positives

**TPR** True positive Rate

**VGG** Visual Geometric Group

**VLA** Vertical long-axis

# Chapter 1

# Introduction

## 1.1 Context and Relevance of the Topic

Cardiovascular diseases (CVD) are the leading cause of morbidity and mortality worldwide [15]. Coronary Artery Disease (CAD), the most common form of CVD, is an atherosclerotic disease responsible for approximately one-third of all deaths among people older than 35 years [16]. In this sense, myocardial perfusion imaging (MPI) by single-photon emission computed tomography (SPECT) plays a crucial role in the diagnosis and management of such cardiovascular conditions.

SPECT is an accurate tool for the non-invasive detection of perfusion changes in the cardiac muscle [17]. The quantification of these images typically involves the extraction of quantitative parameters obtained from the rest stress perfusion SPECT processed data. To perform such tasks software packages are commercially available such as QPS$^{TM}$/QGS$^{TM}$ and 4D-MSPECT. However, these systems have some limitations such as spatial blurring, attenuation, scatter, and low-count data, contributing to high-level noise images characterized by Poisson noise and Gaussian blur. Thus, these aspects may tend to introduce bias in the classification, additionally to the time-consuming and user-dependent, leading to significant intra and inter-operator variability [18, 19].

Furthermore, among various procedures and medical imaging modalities that use ionizing radiation, its use in myocardial perfusion scintigraphy alone is responsible for over 22% of the total effective dose from all medical imaging procedures in the United States [20]. Moreover, a study from 2020 [21] concluded that from 2013 to 2017, cardiac procedures in nuclear medicine showed the highest percentage interval relative to the total dose (40-55%) among others, having a

higher prevalence in Portugal compared to the European Union. To address this issue, advanced imaging devices, such as cadmium-zinc-telluride (CZT) SPECT, have been introduced. These systems have a higher photon sensitivity than conventional gamma cameras, which means that lower doses can be used. Another approach to reducing the dose is to use novel reconstruction algorithms based on this new technology. However, the implementation of these methods is associated with higher costs, which can be difficult to implement in many services [22].

In this sense, both the variability classification issues and the dose reduction concerns can impact the true assessment of SPECT-MPI [23, 24]. Recently, with the rise of Artificial Intelligence and Deep Learning (DL) algorithms, new techniques have been applied to several medical imaging tasks such as image synthesis, reconstruction, segmentation, pre-processing, denoising and classification. Some studies have evaluated the effect of reduced administered activity on assessment of function in cardiac SPECT, [23, 24], and some others have been experimenting with the denoising of low-dose images to map those with standard-dose counterparts with Convolutional Neural Networks (CNN) [25], followed by classification [26]. However, to our knowledge, the classification and performance assessment of low-dose studies without any function to map them to a higher-dose study, when compared with standard-dose ones is a topic that has not been yet much explored.

Thus, there is a need to study this topic particularly with DL techniques that have shown potential when applied on these problematic [25, 26].

## 1.2 Problematic and objectives

Considering the context given in the previous section, this dissertation aims to study the DL classification of full-count and low-count MPI studies. This became important to explore due to the time-consuming and user-dependent processes in the classification task, leading to significant intra and inter-operator variability [18, 19]. Additionally, it is not feasible to acquire several low-count MPI images due to the necessity of injecting multiple times a patient, thus increasing their radiation exposure. To surpass this limitation, to obtain low-count images we can also reduce the exposition time due to the image acquisition process following a Poisson distribution, meaning decreasing the time per projection leads to fewer counts per projection. Moreover, re-acquiring images with reduced counts is not a viable option due to extra time it would take and because repeat studies may differ in other ways, for example in the patient

position or changes in the radiopharmaceutical distribution. On top of that, ethics approval and patient consent would be necessary to acquire repeat studies [27]. So, this process can be simulated using the Poisson resampling method, achieving similar count statistics as the real images, as described in the literature [27]. However, its use regarding DL classification tasks has not been yet much explored.

In this sense, this project centres around two major problems:

- Automatic classification of MPI studies to minimize intra and inter-variability.
- Application of synthetic low-count data regarding the DL classification of SPECT-MPI studies.

To possibly answer the questions proposed with these problems, our main objective is to develop an effective image classification framework for standard full-time and low-time MPI. Specifically, aiming to achieve the following objectives:

- Train a DL model with real full-time MPI studies and evaluate the classifier's performance.
- Application of the Poisson subsampling technique to synthesize low-time images from standard real full-time MPI DICOM studies.
- Train a DL model with real full-time MPI studies and synthetic low-time studies of 75%, 50% and 25% and compare its performance with the one trained solely with real full-time images.

## 1.3 Possible relevant contributes

Overall, as we saw in the previous sections, the classification of standard full-time SPECT-MPI images through DL, and associated to the dose reduction issue, is well documented in the literature [18, 28, 29]. Additionally, the data for simulated reduced counts is typically acquired in a list mode, enabling the use of the same image dataset to simulate various scanning durations and/or injected activity [30]. Contrarily, the simulated low-count dataset used in our work was achieved by using the DICOM non-listed data. In this sense, we believe the main contributions of this thesis are:

- Implementation of a DL technique to classify full-time and low-time SPECT MPI images, in order to minimize the inherent intra and inter-operator variability in these processes.

- Application of a down-sample method from non-list mode DICOM data in order to mimic the SPECT-MPI reduction dose.
- Application of the obtained down-sampled data from non-list mode DICOM in the DL classification of SPECT-MPI.

## 1.4 Thesis structure

This thesis in divided in four chapters. In Chapter 1 we give an introduction to this work, presenting its context, problematic, objectives and some possible relevant contributes.

Chapter 2 describes the state of the art around this thesis topic. We start by introducing cardiovascular diseases and the importance of non-invasive diagnostic tests for its detection. Then, we describe the SPECT-MPI modality, detailing also its clinical image processing methods. At the end, we illustrate the field of DL, explaining its use in medical imaging analysis, in particular in nuclear cardiology imaging.

Chapter 3 presents the experimental work regarding the use of transfer learning for MPI classification. Our results and further discussion are also described in this section.

Chapter 4 includes our conclusions and suggestions for future work.

The Appendix section is presented at the end to give further details of the validation of our Poisson resampling method implementation and of the singular evaluation metrics used in the classification task.

# Chapter 2

# State of the Art

## 2.1 Coronary Artery Disease

Cardiovascular diseases (CVD) have collectively remained the leading causes of death worldwide [15]. More than half a billion people around the world continue to be affected by such diseases, which accounted for 20.5 million deaths in 2021 [31].

Coronary Artery disease (CAD), one of the major CVDs, is a cardiovascular disorder occurring due to atherosclerosis or atherosclerotic occlusions of the coronary arteries. The formation of atherosclerotic plaque in the artery results in the obstruction of blood flow leading to a mismatch between myocardial oxygen demand and supply, a phenomenon called ischemia (see Figure 2.1) [32]. In this sense, reversible ischemia occurs when there is a temporary increase in oxygen demand, for example during exercise or pharmaceutical stress testing and it is manifested clinically by angina pectoris. Conversely, irreversible ischemia may be microscopic, for example in cases of unstable angina [33].

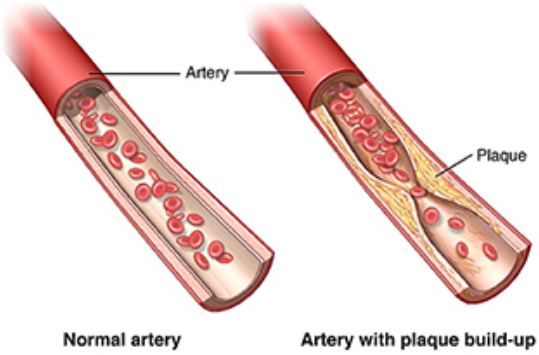



**Figure 2.1:** A normal vessel and an atherosclerotic vessel with clear atheroma build up. Adapted from [1]

There are several risk factors associated with the development of CAD. These are extensively studied in the literature being smoking, diabetes, hyperlipidemia, hypertension, and obesity the major ones reported. [32]

Additionally, the challenge of reducing the burden of CVDs, is particularly relevant in middle-income countries where the disease burden is higher, no doubt reflecting increased rates of hypertension, diabetes, and smoking compared with high-income countries. While the expense of modern cardiovascular technology is hard to prioritize in such countries, steps to tackle and monitor key risk factors may be a realistic strategy for reducing the CVD burden [34].

In this sense, detecting the aetiology of the symptoms, knowing the several risk factors and particularly discerning whether they can be attributed to underlying CAD is essential. In these cases, non-invasive diagnostic tests are recommended to help diagnose and assess the event risk. Thus, Single-photon emission computed tomography (SPECT) is one of the most common functional non-invasive tests that is designed to detect perfusion changes in the cardiac muscle [17].

## 2.2 SPECT

### 2.2.1 SPECT - MPI

Since the early introduction of radioactive tracers for use in medicine, there has been a desire to visualize the true distribution of the tracer and quantify tridimensional (3D) uptake in the body. [35]

SPECT is a non-invasive technique that measures the intensity of gamma rays emitted by small-volume elements. In this sense, in myocardial perfusion imaging (MPI), a radiopharmaceutical is administered intravenously into a patient and partially fixed in the myocardium proportionally to the myocardial blood flow (MBF) and determined by the intrinsic washout. Thus, SPECT measures the three-dimensional concentration of the administered radiopharmaceutical. Additionally, this three-dimensional mapping is composed of various bidimensional planar images (mathematically assigned as projections) acquired by rotating detectors around the object at equally angular interval spaces. Thus, providing true 3D imaging of the organ of interest, while eliminating overlying and underlying source activities. [36]

During a SPECT scan, for each decay, one or more gamma photons ($\gamma$) are emitted with

isotope-specific energies (typically between 100-600 keV) in a random spatial direction. [37]

Based on this principle, SPECT-MPI studies the visualization of the heterogeneity of regional myocardial blood flow (rMBF) secondary to impaired regional coronary flow reserve (CFR) downstream of coronary arteries with significant stenosis. Thus, areas showing less perfusion indicate limited MBF due to stenosis or occlusion, which can be classified as reversible or irreversible based on the rest and stress scan. This means that if it is possible to visualize an area in which the relative perfusion in stress is lower than that in rest it indicates a reversible effect or an ischemia of the vessel. If perfusion in this area is absent both in stress and rest studies, it's indicative of an irreversible occlusion of the vessel, an infarct. In this sense, the comparison between the rest and stress scan it's indicative of the presence or absence of stenosis or occlusion [38].

Most SPECT systems in MPI consist of a dual-head SPECT scanner attached to a gantry usually configured at 90º with parallel-hole collimators, typically equipped with scintillators like thallium-activated sodium iodide (NaI(TI)) and photomultiplier tubes (see Figure 2.2 (A)) . This configuration allows acquisition with an orbit of 180º (right anterior oblique (RAO) to left posterior oblique (LPO) or LPO to RAO) with a rotation of only 90º, thus images are acquired as the gantry rotates the detectors in a continuous or step-and-shoot mode [39, 40].

However, conventional SPECT-MPI has some basic limitations, including long image acquisition times, low image resolution and patient radiation. Usually, to balance the binomial resolution–sensitivity (high-resolution images require a reduction in image sensitivity), scanning time is approximately 15–20 minutes for each stress and rest acquisition, resulting in frequent artifacts caused by patient movement [5]. From a physics point of view, these systems also suffer from spatial blurring, attenuation, scatter, and low data count, which contribute to high-level noise images characterized by Poisson noise and Gaussian blur. [41]

In recent years, specialized cardiac SPECTs have been developed to overcome these obstacles, for example, the IQ-SPECT and D-SPECT (see Figure 2.2(B) and (C)). While the first one uses a unique collimation on a conventional dual-detector SPECT maximizing the sensitivity over the heart, the second one uses solid-state detectors made of cadmium-zinc-telluride (CZT) crystals, allowing for better imaging quality and/or reduced dose and/or time in imaging protocols, mainly due to improved sensitivity and spatial resolution. [39, 42]

Nevertheless, knowing this new generation imaging equipment is associated with higher costs, which can be difficult to enforce in many services, new imaging systems and iterative

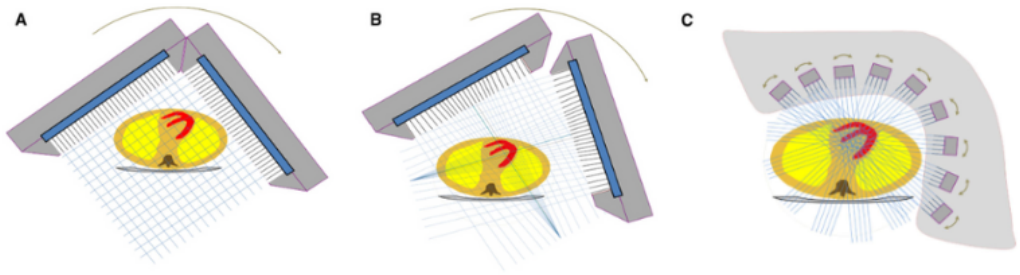



**Figure 2.2:** SPECT system configurations. (**A**) Conventional dual-detector SPECT with parallel-hole collimation. (**B**) IQ-SPECT utilizes the magnifying properties of a focusing collimator near the center of the field of view (FoV), reducing truncation and maximize sensitivity over the heart while avoiding collimation of the body. (**C**) D-SPECT with nine columnar CZT detectors that simultaneously sweep across the heart during acquisition, allowing higher sensitivity and lower dose protocols. Adapted from [2]

reconstruction algorithms have been developed with the aim of increasing signal-to-noise ratio (SNR) and system resolution. Thus, allowing higher photon sensitivity and improving both image quality and resolution [5, 2].

#### 2.2.1.1 Radiopharmaceuticals in SPECT-MPI

Two main types of radiopharmaceuticals are used in MPI: $^{201}$Tl and $^{99m}$Tc-labelled agents such as $^{99m}$Tc-sestamibi and $^{99m}$Tc-tetrofosmin.

$^{201}$Tl was the original and only MPI tracer, but over the years, the nuclear cardiology community has migrated to $^{99m}$Tc-based tracers. $^{201}$Tl emits low-energy photons (71-80 keV) requiring longer imaging acquisition times and resulting in limited image quality due to absorption and photon scattering especially in obese patients. $^{201}$Tl is produced in a cyclotron, having a relatively long half-live (73 hours). In this sense, it requires lower injection doses to minimize radiation exposure. Its biodistribution is generally proportional to the organ blood flow and although a lung uptake is usually low, an increased lung uptake is known to be associated with greater segmental myocardial perfusion abnormality, increased severity and extent of CAD. Additionally, with these radiotracers delayed images are acquired 2 to 4 hours after the initial injection. Through them, it's visible its redistribution from the differential washout rates of normal (faster washout) and ischemic regions (slower washout), giving a sense of normalization [43].

Conversely, $^{99m}$Tc-labeled tracers, such as $^{99m}$Tc-sestamibi and $^{99m}$Tc-tetrofosmin are

currently widely used. Contrary to $^{201}$Tl, $^{99m}$Tc is a generator produced aged eluted from $^{99}$Mo and it has higher energy (140 keV), thus providing better image quality. Its half-life is shorter (6 hours), so the administered dose is relatively larger and the radiation exposure is lower. $^{99m}$Tc-labelled agents are distributed in the myocardium according to rMBF and are excreted via the hepatobiliary system and excreted into the gastrointestinal tract [44]. Due to its slow clearance from the myocytes, separate stress and rest injections are required for the detection of stress-induced ischemia. [43, 22].

In this sense, different radiopharmaceuticals can be used in different clinical situations with their according protocols.

#### 2.2.1.2 SPECT-MPI Protocols

$^{99m}$Tc-labelled myocardial perfusion tracers exhibit minimal redistribution over time, meaning these tracers are taken by the myocytes and are retained within these cells, separate injections are given in order to assess stress and resting perfusions. However, different imaging protocols can be followed: 2-day and 1-day (either stress-rest or rest-stress) [45].

From a dosimetric point of view, a 2-day protocol is preferable. This method consists of acquiring each scan on different days, keeping the total radiation burden to the patient (and staff) at a lower level. Additionally, the stress study, either achieved through previous exercise or pharmacologic methods such as adenosine or dobutamine, should usually be performed first, since the rest study can be omitted if the stress study is interpreted as normal. Regarding the 1-day protocol, where both scans are acquired in a single day, the order of them depends to some extent on the indication for the exam. If the indication is to detect viable myocardium and reversibility of a defect, in a patient with previous infarction, generally is preferable to perform the rest study first. Conversely, if the indication is for diagnosis of myocardial ischaemia, the stress study should be performed first, to avoid the reduction of contrast of a stress-induced defect by a previous normal study. From a dosimetric point of view, this method may be considered sub-optimal due to the necessity of a second tracer injection with three times the activity of the first one [45].

It's important to note that some protocols include dual-isotope imaging, where $^{201}$TI is used for rest imaging and $^{99m}$Tc-labelled tracers for stress imaging, but even so, this method adds radiation burden to patients compared to previously referred protocols, so, nowadays, it's not usually performed. [45]

#### 2.2.1.3 Dose Dosimetry in SPECT-MPI

Radiation exposure is a sensitive topic regarding image modalities that use ionizing radiation. The radiation levels used in clinical settings rarely result in the occurrence of deterministic effects, meaning after exceeding a certain threshold causing effects such as erythema or necrosis. However, cumulative lifetime radiation exposure increases the chance of induction of non-lethal damage to cells, so-called stochastic effects. Theoretically, the chance of the occurrence of a stochastic effect increases with the cumulative radiation dose. For example, the risk of cancer is roughly 0.5% for a cumulative dose of 100 mSv. To provide context, SPECT-MPI averages a radiation dose of 3.0-11.0 mSv per study [46].

In this sense, the medical field is ruled by the ALARA (As Low as Reasonably Achievable) principle, meaning using the lowest possible tracer dose that will still provide diagnostic quality images. Despite the growing radiation awareness in cardiac imaging, the radiation dose is still very high. Thus, studies have shown MPI alone is responsible for over 22% of the total effective dose from all medical imaging procedures in the United States [20]. In Portugal, the SPECT-MPI procedure is more prevalent than in the EU, and it is the one that most contributes to the collective dose in the population since 2010 [21]. Additionally, in 2010, even with the recommendation of the American Society of Nuclear Cardiology (ASNC) that the average radiation exposure for cardiac nuclear imaging studies was 9 mSv in 50% of studies conducted, however, recent studies have already revealed a non-compliance with these practice objectives two years after these recommendations [47, 48].

To meet this purpose, other techniques and recommendations have been developed over time. Regarding the SPECT-MPI protocol, some of them include avoiding dual isotopes due to the patient's maximum radiation dose and performing only stress imaging. That is because if stress studies are normal, subsequent rest studies can be avoided, reducing radiation doses up to 80% [49]. However, similarly, studies have shown non-compliance since only 18% of the centres in the United States and 32.6% in Europe apply such protocol [46]. On an instrumentation approach, camera-based dose reduction strategies (such as highly efficient solid-state SPECT cameras or iterative reconstruction methods) have also been applied. However, this may be associated with higher costs which can be difficult to implement in many services [22, 49].

### 2.2.2 SPECT Clinical Image Processing

#### 2.2.2.1 SPECT Reconstruction

In a tomographic 2D SPECT acquisition, the detectors rotate around the object, which at every angle stores the projection of radioactivity detected in that volume. Thus, in a parallel beam illumination, every projection at angle $\theta$ is defined by the integration along all the parallel line of responses (LOR's) [3]. Before any reconstruction, the projections in the spatial domain are translated into the projection domain through the mathematical representation of the Radon Transform creating a sinogram. Thus, a sinogram represents a grid in which each row of projections represents the sum of all counts along a straight line through the depth of the object being imaged (see Figure 2.3) [50].

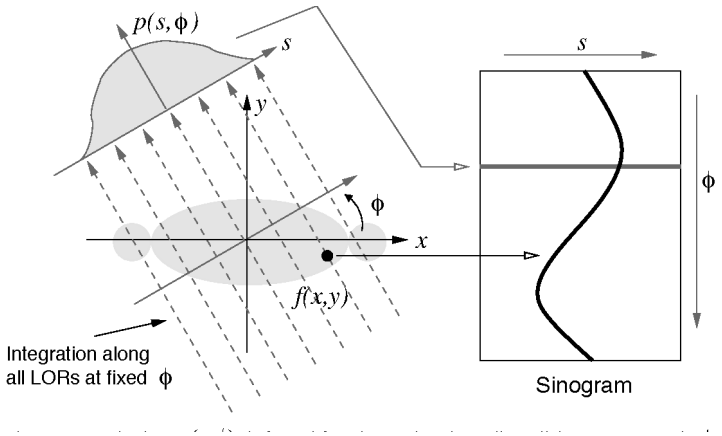



**Figure 2.3:** Projections of the original image and the correspondent sinogram, where every projection at angle $\theta$ is defined by the integration along all the parallel line of responses (LOR's). Adapted from [3]

Image reconstruction is the process of estimating an object image slice $(x, y)$ from a set of projections $p(s, \theta)$ [51]. In nuclear cardiology, image reconstruction has been traditionally performed using analytical methods, being the most common one filtered back-projection (FBP). As stated above, iterative methods are recently also being used [5].

The FBP method is widely used in clinical SPECT because of its simplicity, speed, and computational efficiency. It consists of two steps: filtering the data and back projection of the filtered data.

The simple back projection technique, starting from the sinogram, redistributes the number of counts of each particular point back along a line from which they were originally detected. This phenomenon results in the creation of a star artifact and the blurring of the image (see Figure 2.4 (A)). To suppress this blurring, each projection is first filtered in the Fourier domain

with the ramp filter prior to the back projection. However, despite the ramp filter suppressing low-frequency blurring, it amplifies the already noisy high-frequency data of the projection. In this sense, a combination of the ramp filter and a low pass (*e.g.*, Butterworth) is commonly used, in the frequency domain, to filter the projection data prior to back-projected onto the image matrix, in the spatial domain (see Figure 2.4 (B)) [52, 53, 54].

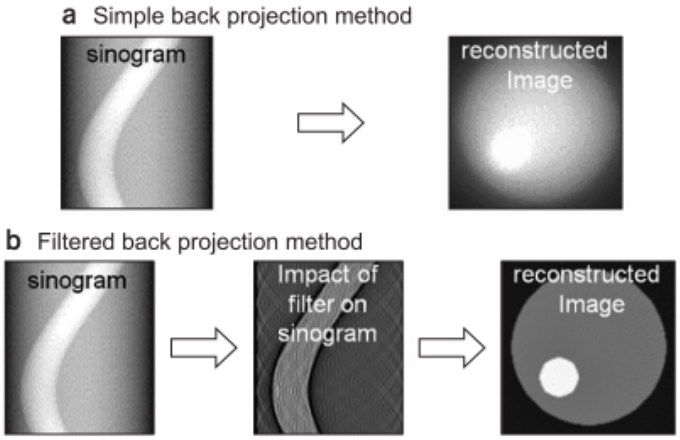



**Figure 2.4:** Computed tomography reconstruction methods. **(A)** Simple back projection algorithm resulting in star artefacts and blurring of the image. **(B)** Filtered back projection algorithm where the sinogram is first filtered before being back-projected onto the image matrix. Adapted from [4]

Regarding iterative methods, as the name implies, the principle of this algorithm is to find the solution by successive estimates. This process is initiated by arbitrarily creating a first estimate. Then, the projections corresponding to the current estimate are compared with the measured projections and the result of the comparison is used to modify the current estimate, thereby creating a new one. There are different reconstruction techniques based on the iterative methods, they differ in the way the measured and estimated projections are compared and the kind of correlation applied to the current estimate [41]. Iterative methods have the advantage of including attenuation and scatter corrections as well as collimator response and more realistic statistical noise models [5].

Iterative algorithms can be classified into two classes: conventional iterative algebraic methods and iterative statistical reconstruction methods.

Examples of conventional iterative algebraic methods are the Algebraic Reconstruction Technique (ART) and the Simultaneous Reconstruction Technique (SIRT). These methods reconstruct the images by solving a set of linear equations. However, due to the large dimensions of the projection matrix, they arise some computational concerns [54, 55].

Regarding iterative statistical reconstruction methods, Maximum-Likelihood Expectation Maximization (MLEM) and Ordered-Subsets Expectation Maximization (OSEM) are some examples of them and their principle is to reconstruct images by assuming Poisson noise and iteratively maximizing a likelihood function. Despite the MLEM algorithm incorporating image degrading effects into the projection matrix, such as scatter, attenuation, depth-dependent resolution and geometrical weighting, thus improving the reconstructed image, this algorithm is too slow for daily routine [54]. To overcome this limitation, Hudson and Larkink introduced the OSEM algorithm in 1994, and since then it become the most frequently used iterative reconstruction algorithm in SPECT. This new approach splits the measured dataset into different subsets and uses only one subset for each iteration (instead of requiring all projections in MLEM). Thus, speeding up the algorithm with a factor equal to the number of subsets [54].

#### 2.2.2.2 Image Reorientation and Display

After the image reconstruction process, follows the critical phase of myocardial processing: reorientation of tomographic data into natural approximate symmetry axes of the patient's heart.

This step can be either performed manually or automatically and results in sectioning the images into vertical long-axis (VLA), horizontal long-axis (HLA), and short-axis (SA) planes of the heart. It is important to note that this step must be consistent and reproducible between rest and stress studies. Inappropriate reorientation can result in misaligned myocardial walls between rest and stress studies, which can potentially result in artifacts and further incorrect interpretation (see Figure 2.5) [5].

The conventional slice display of cardiac rest-stress studies is SA, VLA, HLA. The short-axis images, slices perpendicular to the long axis of the Left Ventricle (LV), should be displayed with the apical slices to the left and the base at the right. The VLA images, slices parallel to the septum, should be displayed with septal slices on the left and lateral slices on the right. The HLA images, slices parallel to the inferior wall, should be displayed with inferior slices on the left and anterior slices on the right. For purposes of comparison of sequential images, they should be displayed aligned and adjacent to each other, with stress followed by rest perfusion, either in rows or columns (see Figure 2.6) [5, 40].

Regarding the image display, different colour maps can be used depending on the operator's expertise, but the use of a linear or a grey colour scale is usually recommended. For the same reason as before, for comparison purposes, the same colour map should be used for rest and stress

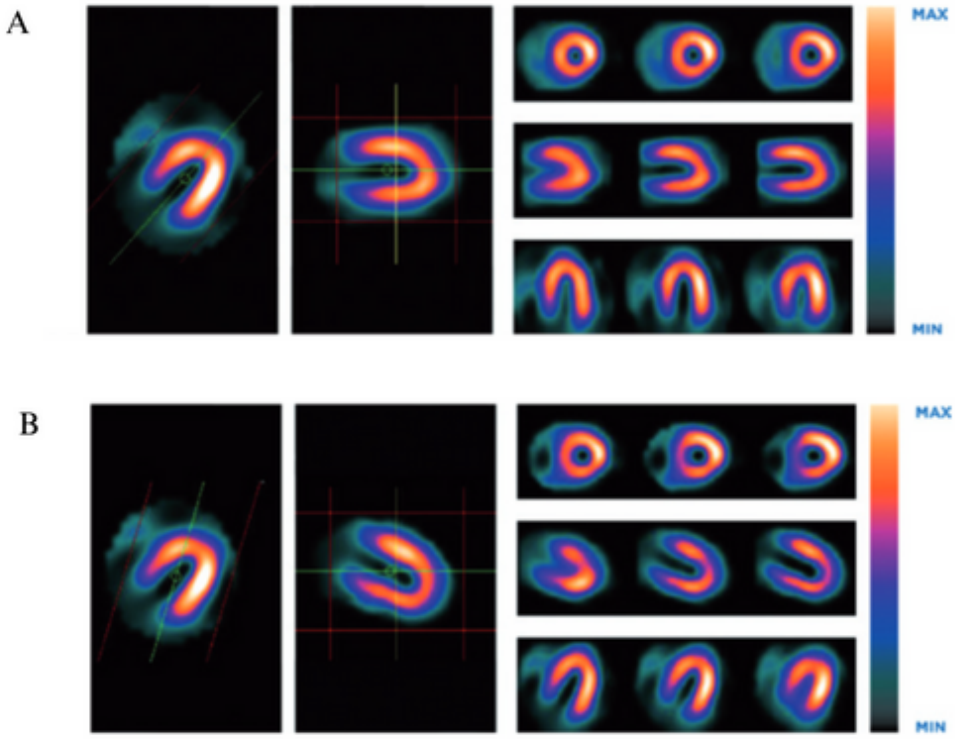



**Figure 2.5:** Differences between a correct and an incorrect reorientation and adjustment of the myocardial walls. (**A**) Example of a correct reorientation and alignment adjustment of the myocardial walls. (**B**) Example of an incorrect reorientation and alignment adjustment of the myocardial walls, potentially resulting in artefacts and further incorrect interpretation. Adapted from [5]

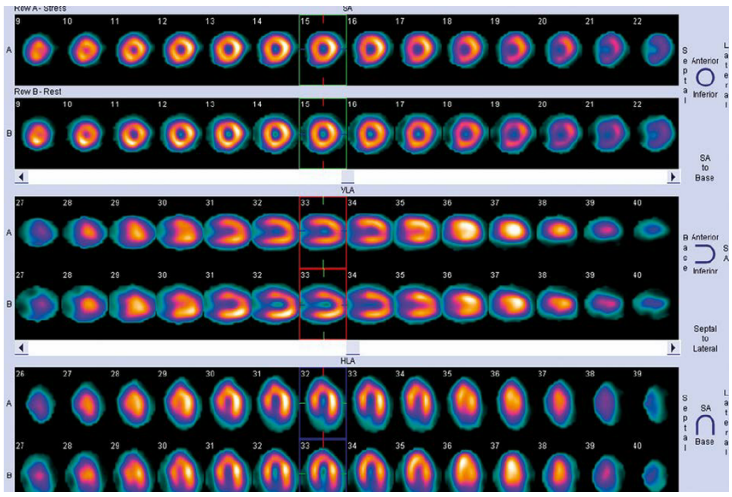


**Figure 2.6:** Standard display of the cardiac slices in a SPECT MPI study. Adapted from [6]

studies. Additionally, there are two commonly used approaches for image normalization, each one with their pros and cons. The first approach is "series normalization" where each cardiac plaque (SA, VLA, HLA), is normalized to the brightest pixel in the entire image set. Although

this is the recommended method to evaluate the extent and severity of perfusion defects, this approach is sensitive to focal hot spots, tending to have poor visualization of normal structures at the base and apex of the LV, and lacks an ideal display of each individual slice. The second approach is called "frame normalization" where it finds the the brightest pixel of each slice and normalizes the image accordingly. This approach avoids attenuation artifacts, particularly in the mid and basal segments of the myocardium, and has been shown to result in both high sensitivity and specificity for CAD detection [40].

#### 2.2.2.3 Image Quantification

After the data reconstruction, cardiac plane orientation and proper image normalization follow image quantification. This is a valuable step in MPI involving the extraction of quantitative parameters derivable from the rest-stress perfusion SPECT study. In order to extract the various quantitative parameters, the first step is to segment the LV. This consists of the separation of the region or structure of interest from the remainder of an image [5].

Selected our area of interest, it is possible to obtain the perfusion parameters such as defect extent severity values with polar map display, reversibility stress-rest, percentage of hypoperfused myocardium, percentage of ischaemic myocardium, total perfusion deficit and other categorical summed scores based on the difference between the stress and rest studies (summed stress score (SSS), summed rest score (SRS) and summed difference score (SDS)). Additionally, global functional parameters of the LV can be obtained such as left ventricle ejection fraction (LVEF), end-diastolic volume (EDV) and end-systolic volume (ESV), in conjunction with regional parameters like regional wall motion and thickening of the myocardial wall [5].

To extract these LV perfusion values and functional parameters, there are several software commercially available. Among the most popular used is QPS$^{TM}$/QGS$^{TM}$, developed at Cedars-Sinai Medical Centre (Los Angeles, California) (see Figure 2.7).

Moreover, recently, machine learning and deep learning algorithms have been applied and vastly studied in areas such as image reconstruction, segmentation and pre-processing, denoising, classification and other cardiac imaging analysis. In this sense, these new technologies have been shown to improve diagnosis and prognostication for patients with cardiac disease [56].

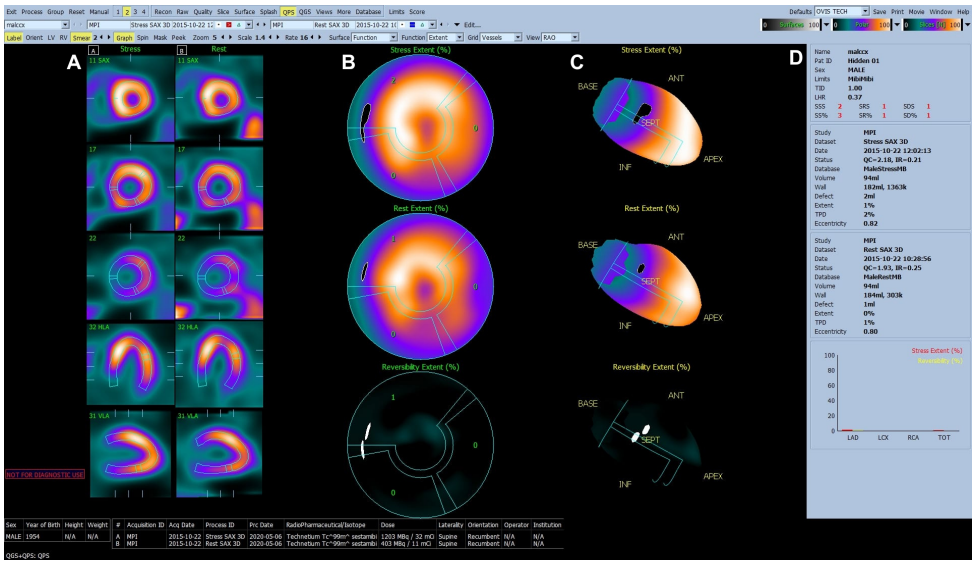



**Figure 2.7:** Cedars cardiac processing working stations. (**A**) Display of the LV in the SA, HLA and VLA axis for both stress and rest studies. (**B**) Polar maps of the stress/rest regions showing radiopharmaceutical uptake throughout the LV and a polar maps identifying any reversible perfusion between both acquisitions. (**C**) Radiopharmaceutical uptake of the stress/rest studies in the shape of the LV. (**D**) Studies characteristics and LV quantitative parameters derived from stress/rest studies. Adapted from [7]

## 2.3 Artificial Intelligence

In recent years, with the advancements in algorithmic theory, computing power and the availability of large amounts of data, we have been experiencing a rise in Artificial Intelligence (AI) algorithms for medical imaging applications [57].

The term AI refers to computers that perform tasks that usually require human intelligence. Machine Learning (ML), a branch of AI, refers to algorithms that are programmed to learn from observations and make statistical inferences based on what they learned. For example, suppose a ML algorithm is trained on a set of data (e.g., myocardial perfusion images) and has some knowledge of the data (*e.g.*, normal and ischemic data). In that case, the model can learn from that data and predict whether the new image is normal or ischemic [57, 58].

The example above is just one of many applications a ML model can have. In general, according to Zemouri [59], ML models in biomedical applications can be used in three main orientations: (1) as a computer-aided diagnosis to help the physicians for an efficient and early diagnosis, with a better harmonization and less contradictory diagnosis; (2) to enhance the medical care of patients with better-personalized therapies; and (3) to improve the human well-being. Nowadays, the leading ML tool in the biomedical field is centered on Deep Learning (DL) with the use of Artificial Neural Networks (ANN) [59].

### 2.3.1 Deep Learning and Artificial Neural Networks

DL is a branch of ML that combines computer science, statistics, and decision theory to find patterns in complex data, often in large quantities (see Figure 2.8). As mentioned before, DL techniques have recently revolutionized many fields, from computer vision and natural image processing to speech recognition and language processing. The advantages of these techniques are obtaining models more robust, generalizable to different types of data or applications and scalable [60]. The algorithms behind these breakthroughs are composed mainly of artificial neural networks (ANN) [61, 62].

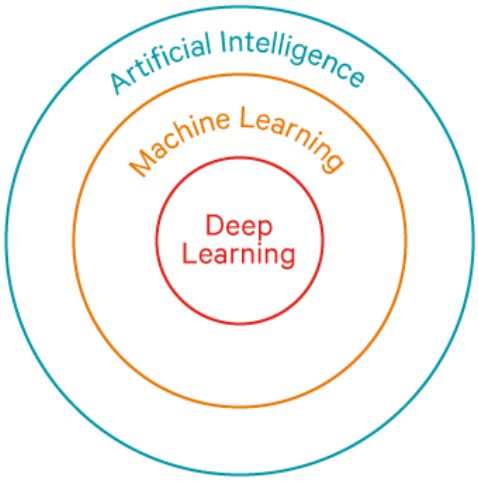



**Figure 2.8:** Relationship between Artificial Intelligence, Machine Learning and Deep Learning. Adapted from [8]

ANNs were inspired in the 1960s by the biological brain neuron networks. They are composed of layers of interconnected units (neurons), and from a mathematical point of view, they are composed of the nonlinear transformation $y = F(x)$ of the input $x$. Between two interconnected neurons there is a weight $w_{ji}$ associated with the connection, which is updated during the learning phase based on the backpropagation of the current total loss. The input data is distributed from the input layer to the output layer, neuron after neuron. This propagation transforms these data from one space to another by the neurons of the layer in a nonlinear way. Each neuron calculates the weighted sum of its inputs and uses a nonlinear activation function (serving as a threshold) to calculate its output. Based on the hidden layers (middle layers) an ANN can be shallow (if they are composed of 1 or 2), or deep (if they are composed of more). In practice nowadays, and since the generalization problem often requires large amounts of data, deep neural networks (DNN) are almost exclusively used as they offer a much higher capacity to learn fitting and feature extracting from high-complexity input data [59]. In Figure 2.9, is

illustrated a feed-forward multi-layer perceptron architecture with a single hidden layer, as an example.

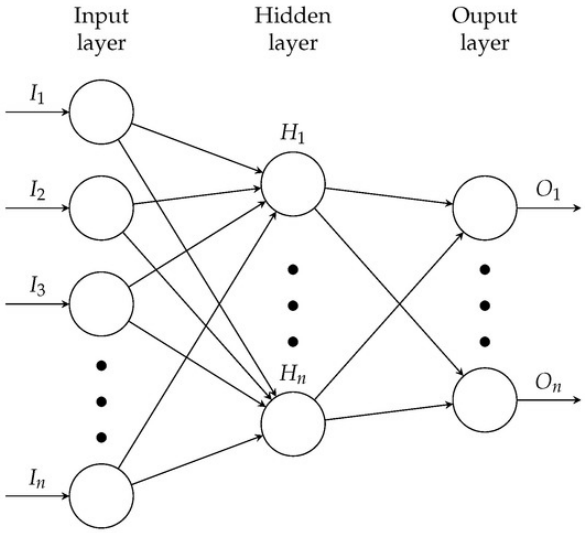



**Figure 2.9:** Example of a multi-layer perception architecture with a single hidden layer. Adapted from [9]

#### 2.3.1.1 Types of Learning

It is important to note that there are several kinds of learning algorithms in neural networks (NNs). Each of them has its advantages and disadvantages and is used for specific situations and problems. In this sense, supervised learning, unsupervised learning, and self-supervised learning are assumed in the literature.

- Supervised learning: In this technique, the NN during the training phase has access to both the input data and its label to learn insights for a target problem. These settings are most commonly used in image classification tasks.

- Unsupervised learning: In this technique, the NN does not have the labels of the training data (unlabeled data). In this sense, the NN has the goal of identifying hidden patterns or the natural structure of the data.

- Self-supervised learning: This method can be considered a special case of unsupervised learning because it also uses unlabeled data. This technique offers promising solutions to the challenges of collecting large-scale solutions, as it creates generalist models that can be fine-tuned for many downstream tasks without large labelled datasets [63].

As there are many types of learning, there are also several types of DL networks that are highly studied in the literature, each one with a specific purpose for each specific task. Thus, convolutional neural networks (CNNs) have been applied to several medical imaging tasks. In the next subsection, this DL network will be more deeply described.

### 2.3.2 Convolutional Neural Networks

In DL, CNN is the most famous and widely used architecture. In addition to what was previously said, CNN has been applied in several areas such as computer vision, speech recognition, face recognition, and others [60]. In particular, in medical imaging applications, they are widely used for disease classification, object detection, and object segmentation [64].

#### 2.3.2.1 CNN overall Architecture

CNNs consist mainly of four types of layers. These are input layers, convolutional layers, pooling layers and fully connected layers (see Figure 2.10). When these layers are stacked, a basic CNN architecture is created. In a classification task, the output of the fully connected layer is a translation of the input images into categories with labels [65].

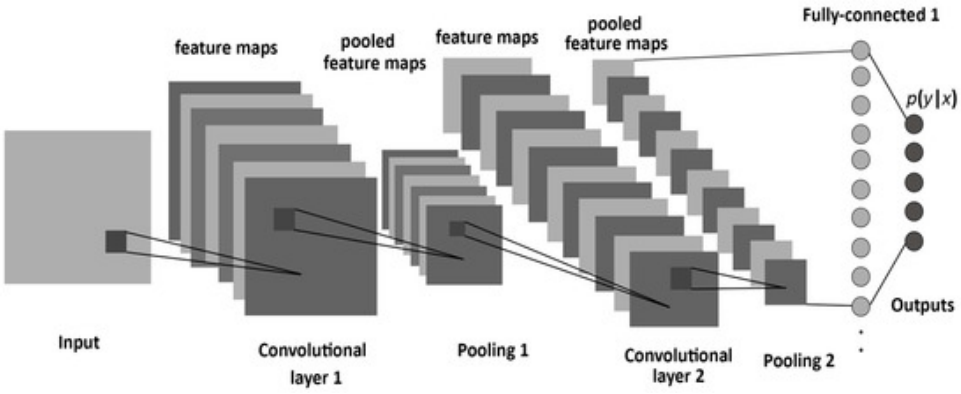



**Figure 2.10:** Typical architecture of a CNN consisting of an input layer, convolutional layers, pooling layers, and fully connected layers at the end responsible for giving the final CNN output. Adapted from [10]

- Input layer - The first layer of each CNN. The input layer takes the input images and resizes them to pass them onto further layers for feature extraction.
- Convolutional layers - The convolution operation corresponds to the application of a convolutional filter over the whole input data. Convolutional filters (also called kernels) consist of a grid of discrete numbers or values, and each value is called the weight of the

kernel. At the beginning of the NN training, these weights are tipically randomly assigned, but after each iteration of training, they are updated to better extract important features and create the so-called feature maps. Analytically, this operation consists of a dot product between a sliding window from the image and the kernel, where their corresponding values are multiplied and then summed up to create a single scalar value as illustrated in Figure 2.11. This process is then repeated until no further sliding is possible.

- Pooling layers - The purpose of these layers is simply to perform the downsampling of the feature maps. In practice, this approach reduces the large-size feature maps to create smaller feature maps while maintaining the majority of the dominant information in every step of the pooling stage. The most common aggregation choices are taking the average or the maximum of the values, as the average pooling method or the max pooling method, respectively, as illustrated in Figure 2.12.

- Fully-connected layers - These are the last layer of every CNN architecture. In these layers, each neuron is connected to all neurons of the previous layer as standard ANNs as previously discussed. Their input is in the form of a vector (1D), which is created from the feature maps after flattening and they output the class scores representing the final CNN output.

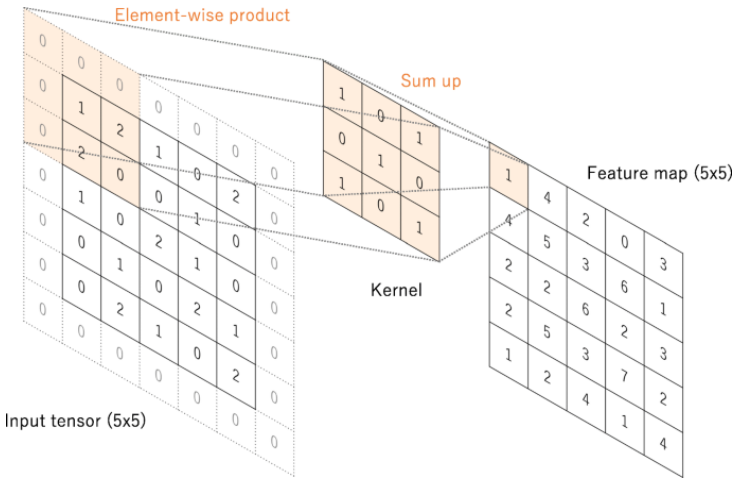



**Figure 2.11:** Example of the convolution operation, where a 3x3 kernel is applied on a two-dimensional input. The resulting output is referred to as a feature map. Adapted from [11]

Other key components of CNN are activation functions and loss functions. Similar to simpler ANNs, as mentioned above, activation functions are used to provide nonlinearity to the algorithm's learning phase so that they can focus on relevant data properties and improve

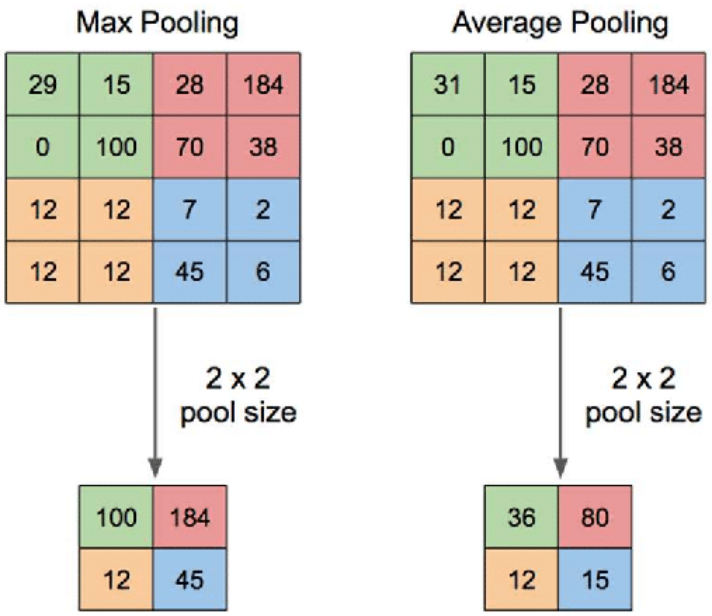



**Figure 2.12:** Example of the average pooling and max pooling operations. Adapted from [12]

performance in solving complex problems. Some of the most used activation functions in CNNs are the Rectified Linear Unit (ReLu) and Sigmoid. On the other hand, loss functions compare target values and predicted output values and measure the performance of NN model training data. The most used loss function in CNNs for classification tasks is the Cross Entropy. Thus, during the learning phase, the main objective is to minimize this loss in each iteration, which means better data fitting. Towards this goal, an optimizer is also defined to adjust the model parameters, being the three most common ones the Gradient Descent, the Stochastic Gradient Descent (SGD) and the Adaptive Moment Estimation (Adam).

#### 2.3.2.2 ResNet Architecture

He et al. [66] created the Deep Residual Network (also known as ResNet) in 2016. This architecture was developed to bypass some of the limitations of DL, which included having a specific number of layers and being time-consuming. In this sense, this network is built on the residual training framework, which boosts the effectiveness of deep network training by using skip connections or shortcuts. Their advantages over the others are that their performance does not suffer as a result of deeper architecture, and the ability to train networks is improved due to lighter computations [67].

The residual block is defined by Equation 2.1:

$$y = F(x, W + x) \tag{2.1}$$

Where, $x$ is the input layer, $y$ is the output layer, and $F$ function is represented by the residual map.

Additionally, these residual blocks on ResNet can only be accomplished if the input data dimensions are identical to the output data dimensions [67].

At the time, several types of ResNet were developed based on the number of layers. To give some examples, we have ResNet-18, ResNet-110, ResNet-1202 etc. With the increase in the number of layers, there is a subsequent increase in the number of parameters. In 2015, ResNet was the winner of the 2015 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) with 152 layers of depth model [60]. In Figure 2.13 is illustrated the architecture of ResNet-18.

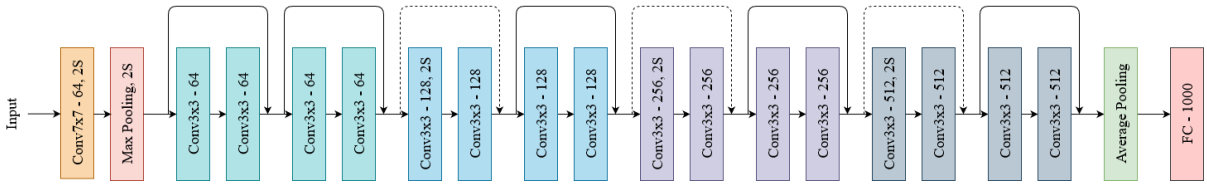



**Figure 2.13:** ResNet-18 architecture

#### 2.3.2.3 Transfer Learning

DL algorithms requires large amounts of data for training because they have many parameters to adapt. This problem is more critical in medical imaging since the limited size of cohorts and the cost of expert annotated datasets, make data scarcity a problem. To overcome this limitation, many researchers use transfer learning techniques. This aims to achieve high performance on target tasks by leveraging knowledge learning from other tasks [68].

Regarding CNNs, there are several pre-trained models publicly accessible with downloadable parameters. The most commonly used were trained on the ImageNet dataset (containing 1,281,167 training images distributed in 1000 object classes). Some available models are, for example, the AlexNet [69], GoogLeNet [70], VGG [71] and ResNet.

### 2.3.3 Explainable AI methods for CNNs

As we already saw CNNs are achieving significant results in the field of computer vision performing several kinds of tasks such as image classification, object detection and segmentation.

However, despite these achievements, the sophisticated structure and complexity of CNNs limit the ability to understand the reasons behind their decisions. Thus, over the years, there has been an increasing demand for CNNs explainability [72].

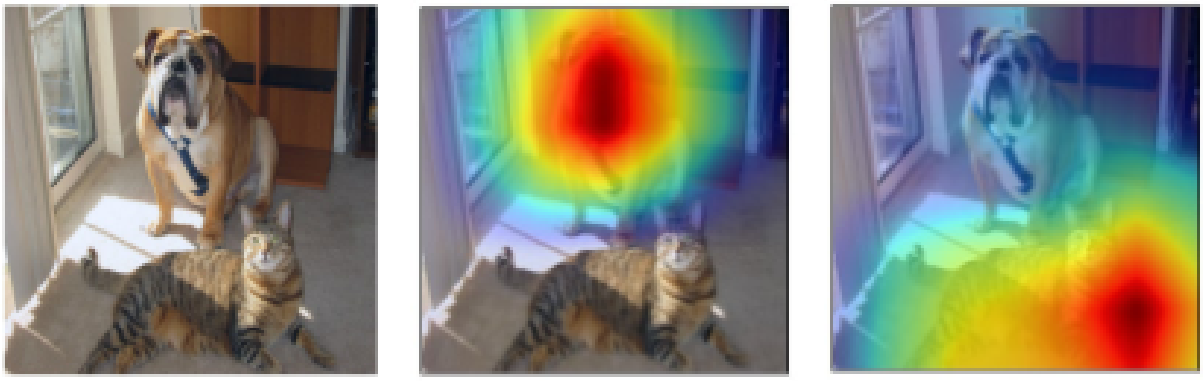
In this sense, several explainable AI techniques have been developed and used to further give insights into the decision-making process of CNNs. In the literature, these methods are divided into two categories: decision models and architecture models. Decision models correspond to techniques applying backpropagation and mapping the predicted class with corresponding pixels in the input image, identifying parts of the image that mostly contributed to the network decision. Meanwhile, architecture models explore the network and analyze the mechanism of layers and neurons [72].

For brevity purposes, only a visual explanation technique subcategorized in the decision models will be detailed, more specifically the gradient-weighted class activation mapping, since we used this method in our experimental work.

#### 2.3.3.1 Gradient-weighted Class Activation Mapping (Grad-CAM)

Grad-CAM is an optimized version of the class activation map (CAM). CAM was proposed in 2015 to interpret CNN predictions by generating activation maps [73]. This method modifies the CNN architecture by adding a global average pooling layer (GAP) instead of the usual fully connected layer. In this sense, the GAP layer computes the average contribution of each feature map in the last convolutional layer, thus generating an activation map overlaying the input image identifying the areas that most contributed to its prediction. However, the drawback of this method is the architecture modification which impacted the prediction accuracy. To surpass this limitation, Grad-CAM was proposed a year later [72, 13].

The main difference between Grad-CAM and the previous model is that this one maintains the fully connected layer and calculates the gradients of the predicted class in the last convolutional layer, thus preserving the prediction accuracy. This model is class-specific to the feature map generated, in this sense it captures features that positively influence the class prediction and ignores negative features irrelevant to the class as illustrated in Figure 2.14 in a cat/dog binary classification. However, despite the literature showing that this model outperformed other gradient-based models it has the limitation of not capturing multiple objects of the same class. Afterwards, this limitation was further worked on with the implementation of optimized techniques such as Grad-CAM++ [72].



(a) (b) (c)

**Figure 2.14:** Grad-CAM analysis for a Cat/Dog image classification task. **(a)** Original image.**(b)** Grad-CAM analysis for the class "Dog". **(c)** Grad-CAM analysis for the class "Cat". Adapted from [13]

## 2.4 Review of Image Processing Techniques and Deep Learning Approaches in Medical Imaging

### 2.4.1 Application of Deep Learning in Medical Imaging

Medical image analysis is a robust research topic, with millions of studies published in recent decades. Recently, the increase in DL algorithms has led to a shift towards non-handmade engineering that allows automated image analysis (see Figure 2.15) [68]. Additionally, these developments are not limited to the research setting, nowadays, DL-based algorithms and software are making their way into clinical environments with their approval by the Food and Drug Administration (FDA) [62].

Hence, image synthesis is one of the main tasks for DL techniques, possibly to help gather data for further model training. In this sense, generative adversarial networks (GANs) have been widely used for it, generating high-quality images over several modalities such as MRI [74] and CT [75]. Additionally, on this topic, GANs have also shown the ability to perform cross-modality image synthesis, such as generating CT images from MRI ones [76].

Additionally, CNNs have been at the core of medical image analysis [77]. For example, CNNs have shown the capability to reconstruct SPECT images with a limited number of projections outperforming methods such as MLEM [78]. For instance, CNNs also have been used in segmentation tasks in CT [79] and MRI [80].

Addressing the physics limitations in the acquisition count statistics in SPECT, several architectures of DL models have been developed to denoising tasks. For example, in a study

from 2023, they compared four different DL architectures, a CNN, a residual network, a UNET, and conditional GAN (cGAN), and denoised full-time, half-time and quarter-time cardiac SPECT images. Thus, the coefficient of variation of myocardium counts was on average 7% (CNN), 8% (residual network), 7% (UNET) and 14% (cGAN) lower than with the iterative reconstruction method OSEM, concluding that DL-based methods effectively improved noise levels [81].

Regarding the classification task, several different architectures of the CNN have been vastly used. Some examples are the deep residual model ResNet-18 used to classify thyroid tissue residues in patients with thyroid cancer after ablation [82] and the use of the VGG-based (visual geometric group) models to classify whole-body SPECT bone scan images [83]. In the classification of SPECT-MPI, extensive work has been also developed as we are going to see in the next subsection.

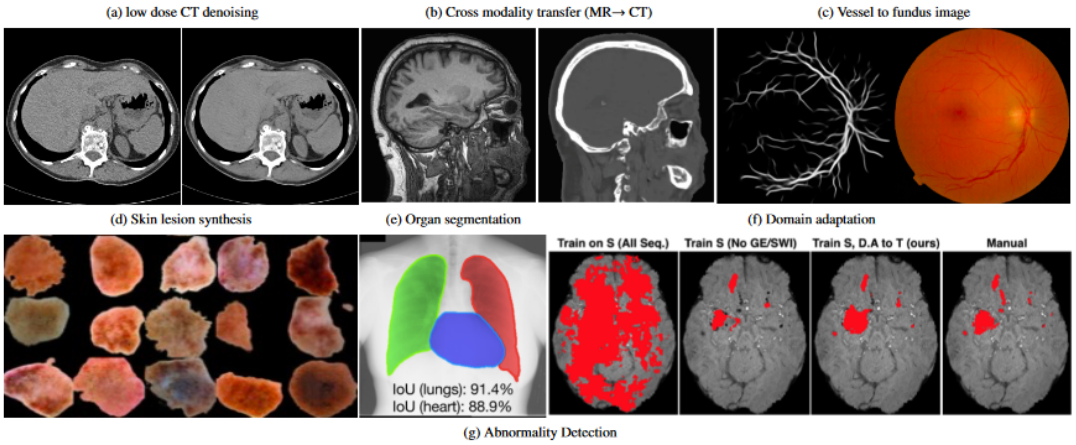



**Figure 2.15:** Deep Learning applications in medical imaging analysis. Adapted from [14]

### 2.4.2 Image Processing and Deep Learning Approaches for Myocardial Perfusion Imaging Classification

Before DL was at the core of classification tasks, several image-processing tools with conventional ML algorithms were and still are widely used. The process of extracting a very large set of quantitative features describing for example the shape and texture of images is called radiomics [56]. Several studies have been using this technique in medical images, and more specifically for MPI classification, and creating big-data databases for ML analysis. For example, in a 2021 paper [84], they applied radiomics for the auxiliary diagnosis of myocardial ischemia using left ventricle (LV) tomograms obtained from D-SPECT MPI. Thus, with the radiomics model obtaining an AUC above 0.95 in identifying the presence or absence of

myocardial ischemia, they concluded this method has good application efficacy and high application potential. In another study, in 2013, Arsanjani et al. [85] trained a boosted ensemble ML model with quantitative perfusion data and clinical data to predict obstructive CAD $\geq$ 70% and concluded the model performs with a diagnostic accuracy of the level of an experts analysis.

In recent years, because of the large number of cardiac images that are routinely acquired with a wide range of modalities and access to public datasets, there has been a surge in publications applying DL in the cardiac domain [62]. In medicine, they have been successfully applied to automate myocardial perfusion imaging interpretation. S. Kaplan Berkaya et al. [18] trained two different classification models, one DL-based and another expert knowledge-based to identify myocardial perfusion abnormalities (ischemia and/or infarction) and they concluded that the DL model performs close to the level of expert analysis. In another study, they trained a NN to detect CAD using myocardial perfusion images from 135 patients and it showed that the NN model got higher sensitivity than the blind experts' classification (0.92 vs 0.71) [86]. Guner et al. [28] developed an open-source framework of NN for the diagnosis of CAD from myocardial perfusion images from SPECT. They trained a NN model over 243 images and compared its results with an expert's evaluation and the results showed there were no statistical differences between the two. Thus, several studies have shown that applying neural networks as clinical decision support tools appears to have significant potential in the future [87, 86].

# Chapter 3

# Experimental Work

## 3.1 Data and Protocols

This study included a cohort of 129 patients clinically referred for SPECT-MPI imaging, at the NuclearMed, Hospital Particular de Almada (HPA). All studies were acquired between December 2021 and February 2023. From those, in addition to the full-count acquisition, 21 studies were acquired proportionally to 75%, 50% and 25% time reduction of the conventional 100% full-time SPECT-MPI (further details are represented in subsection 3.5.1). This single-institution study was approved by the ESTeSL-IPL ethics committee (CE-ESTeSL-Nº26-2022), and all patients gave written informed consent, as the model implemented in this department. The inclusion criteria incorporate all the patients referred for SPECT-MPI with or suspected CAD. Related to the collected sample, the data characteristics are represented in Table 3.1.

**Table 3.1:** Patients characteristics. *min - minimum; Max - maximum; SD - standard deviation*

| | | n | % | min | Max | mean ± SD |
|---|---|---|---|---|---|---|
| Sex | Male | 68 | 52.71 | | | |
| | Female | 61 | 47.29 | | | |
| Age | | | | 43 | 90 | 70 ± 10 |

All studies were performed using either the 1-day protocol or the 2-day protocol. Considering the 1-day method, the first SPECT study (Rest) involved the administration of approximately 400 MBq of $^{99m}$Tc-tetrofosmin, through imaging was initiated 1h post-injection. The second SPECT acquisition (Stress) started with a threefold increase in activity (1200 MBq), initialized

after 1.5h. The 2-day protocol started with the stress SPECT study, and the rest phase was performed on a different day. The administered activities applied in this protocol were similar (600 MBq). Considering both protocols, the stress test was performed using an infusion of adenosine, dobutamine or regadenasone and the administered activities were adjusted according to the patient's weight.

## 3.2 Data Acquisition

A full-time conventional non-gated SPECT-MPI imaging was performed using an Infinia® gamma-camera model (General Electric, GE), equipped with a thallium-activated Sodium Iodide (NaI(Tl)) crystal. The data was acquired in accordance with the acquisition parameters protocol implemented at the Hospital such that the following were used: low energy high resolution (LEHR) collimators, energy window centred on the 140 keV photopic (±10%), 64x64 pixels matrix, L-mode detectors geometry (positioned on the thorax), 180°of rotation from 45°right anterior oblique (RAO) to the 45°left posterior oblique (LPO) view, step and shoot mode, 3°angular interval and a total of 60 projections. In the full-time acquisition, the time per projection was adjusted according to the count rate in the left ventricle so that a total of 1.5 million counts were collected. Additionally, a sub-sampled of 21 patients were sequentially acquired using the same parameters were performed, with a full-time reduction of 75%, 50% and 25%. In all acquisitions, the patient's position was preserved and consistent, as in dorsal decubitus, feet-first, and with arms elevated above the head, sustained by an arms support.

Previously to image reconstruction, all SPECT-MPI acquisitions were technically inspected to evaluate the presence of possible artifacts via visual assessment of the sinogram. Also, the non-presence of subdiaphragmatic tracer concentration related to the radiopharmaceutical biodistribution was assured.

## 3.3 Data Reconstruction

All acquisitions were reconstructed using the Xeleris 2.0 clinical workstation (General Electric, GE). The analytical method of Filtered Back Projection (FPB) was used to reconstruct all studies for clinical purposes. To balance the images' signal-to-noise ratio (SNR), the Wiener filter (FBP-W) was also used, based on Equation 3.1. The reconstruction and filtering methods of

choice followed the standard method used by the HPA nuclear medicine department physicians.

$$W(f) = MTF^{-1} \times \frac{MTF^2}{MTF^2 + Pn/PF} \tag{3.1}$$

where W(f) is the value of the wiener filter as a function of spatial frequency, MTF is the modulation transfer function of the imaging system, Pn is the noise power spectrum, and PF is the signal power spectrum. The quotient Pn/PF can be interpreted as signal-to-noise ratio (SNR).

The clinical implementation of the FBP-W algorithm was performed applying a Point Spread of 3 cycles.pixel$^{-1}$ and a Pn/PF of 0.25 cycles.pixel$^{-1}$.

For each patient, considering all the raw-data acquisitions, the reconstruction volume and the slice thickness of 6.35 mm were equal, in order to obtain matching slices inter-acquisition and inter-patient.

## 3.4 Image Resampling

Medical images present an ideal format for DL tasks, however, it is difficult to secure large quantities of data because of patient privacy and security policies, such as the Health Insurance Portability and Accountability Act (HIPAA) privacy rule, and because it is necessary ethics approval and patient consent to acquire repeat studies. [88, 27]. To surpass these limitations, specifically in medical imaging classification tasks, several techniques have been applied such as image augmentation and image synthesis.

One of our objectives is to compare the classification of standard-time and low-time data, knowing it is not feasible to inject several times a patient and also due to the extended time it would take, in order to acquire more data, specifically for the training phase, we can simulate this process through image resampling techniques. Accordingly, the literature refers to the use of image resampling methods such as binomial resampling [29, 89, 90], bootstrap resampling [91] and Poisson resampling [22].

In this work, the stress studies of the whole dataset were resampled using the Poisson resampling method as described in [27] with 25%, 50% and 75% count settings.

### 3.4.1 The Poisson Resampling Method

During the SPECT acquisition process, the noise for counts obeys the Poisson distribution. Thus, in this context, the Poisson resampling method tries to produce simulated low-count images from full-count images considering the noise distribution in the given image.

In Poisson sampling, each element $i$ in a population is included in the sample independently of all other units with a probability $\pi_i$. Generally, in this method, each element may have a different probability of being included in the sample. For purposes of simulating a reduced-count image, each count in the image has the same probability of being selected for the sample. To achieve this goal, the Bernoulli trial is applied.

Starting by defining the probability $\pi_i$ of selection in the new image, for example, 0.5 for a half-count image, for each count of each pixel is allocated an independent uniform random number $U_1 \ldots U_N$ on [0,1] where N is the number of counts in a given pixel. This count is then subjected to the Bernoulli trial to see if it is less than or equal to the probability of selection ($U_c \leq \pi_i$). If true, the count is selected for inclusion to define the pixel value of the new image, otherwise it is discarded. In the end, we get our simulated low-count image with the percentage previously defined (by the probability $\pi_i$).

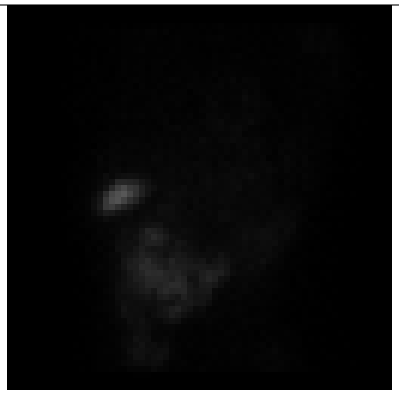
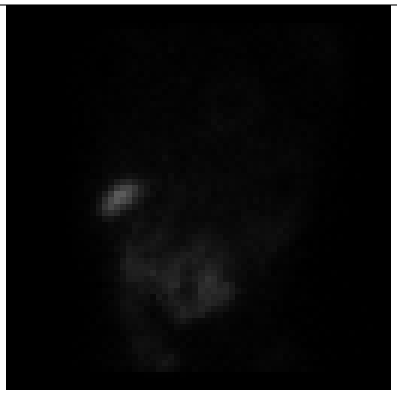
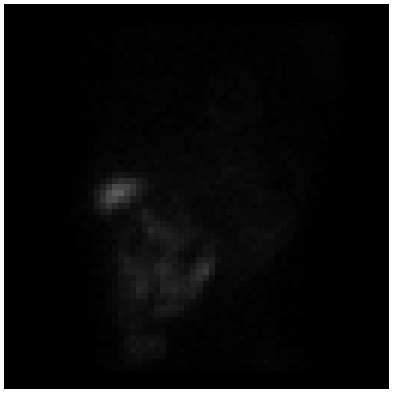
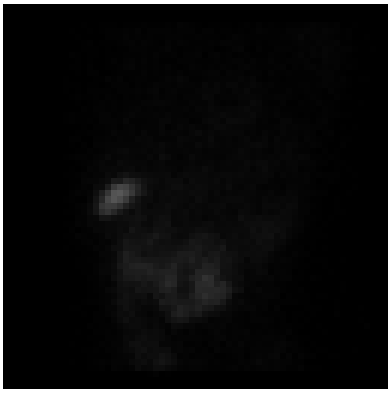
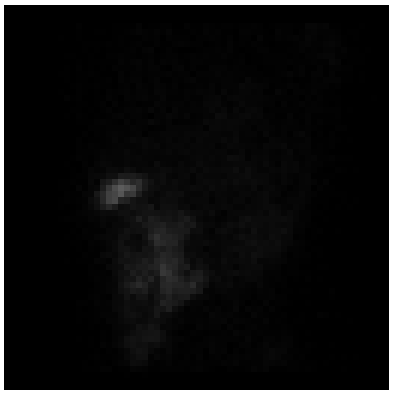
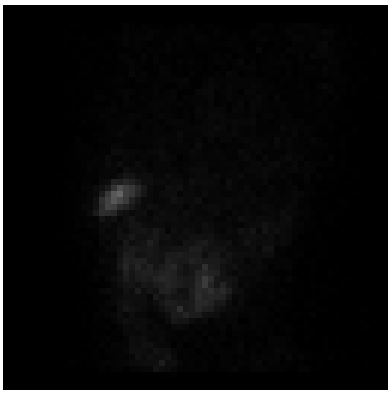
In this sense, we applied this methodology, as described in [27], by the function described in section A.1. As an example, in Table 3.2 we can see the real and the resulting subsampled images in the $30^{\text{th}}$ projection of a study. Despite the well-validated method in the literature, we proceeded to do a brief validation of the method comparing our resulting images to the real low-time ones. Further analysis can be assessed in subsection A.1.1.

## 3.5 Transfer Learning for Myocardial Perfusion Imaging Classification

In this dissertation, we ran 5 training jobs to binary classify myocardial perfusion images between normal and abnormal (from ischemia or infarct): 1 with post-reconstructed SPECT-MPI slices from a stress full-time acquisition (100R), 3 of them with the subsampled images generated from the Poisson resampling method described in section 3.4 with 3 distinct percentages of 75%, 50% and 25%, as 75S, 50S and 25S; and the final one with these 4 datasets combined (ALL).

In this sense, the purpose of the 100R training job is to assess the performance of the model

**Table 3.2:** Real and subsampled 30$^{th}$ projection of the study H134066

| | Real | Poisson Subsampled |
|---|---|---|
| 75% |  |  |
| 50% |  |  |
| 25% |  |  |

to classify normal and abnormal images in different percentages of data counts when trained solely on full-time count data. Then, we ran the 75S, 50S and 25S training jobs to assess the viability of using synthetic data and understand their ability to classify real and synthetic data of the same percentage. Lastly, we ran a final model with all the previous datasets combined and trained to further compare its results with the 100R model.

**Table 3.3:** Test data characteristics. *min - minimum; Max - maximum; SD - standard deviation*

| | | n | % | min | Max | mean ± SD |
|---|---|---|---|---|---|---|
| Sex | Male | 11 | 52,4 | | | |
| | Female | 10 | 47,6 | | | |
| Age | | | | 44 | 79 | 67.0 ± 8.9 |
| Total Counts | 100% | | | 7.4e5 | 9.3e6 | 7.0e6 ± 2.0e6 |
| | 75% | | | 3.0e6 | 6.7e6 | 5.2e6 ± 8.7e5 |
| | 50% | | | 1.7e6 | 4.3e6 | 3.4e6 ± 6.7e5 |
| | 25% | | | 1.3e6 | 2.5e6 | 1.8e6 ± 3.1e5 |

### 3.5.1 Dataset

The dataset described in section 3.1 was split into train, validation and testing sets. The testing set contained the 21 studies referred in section 3.1, since they were the only ones that full-time images (100Rtest) and reduced count ones were acquired (75Rtest, 50Rtest, 25Rtest). The characteristics of this subgroup is represented in Table 3.3. To additionally assess the performance of the classifiers in synthetic data, three low-time synthetic test sets were generated through the Poisson resampling method from the real full-time test set images with a reduced percentage of 25%, 50% and 75%, as 25Stest, 50Stest and 75Stest. The validation set was used to choose the best model and optimise it while tuning the hyperparameters and it consisted of a total of 40 studies (20 "normal" and 20 "abnormal"). The remaining studies, which were sixty-eight (68), became part of the training set. In Table 3.4, we can see the dataset distribution of the train, test and validation sets according to its labels.

### 3.5.2 Data Preparation

All studies were available in DICOM (Digital Imaging and Communications in Medicine) format, with 16-bit depth. First, we extracted every individual slice of each axis (SA, VLA, HLA). Prior to saving each image, we converted them to 8-bits of depth to have a grayscale image ranging between 0-255 and normalized them to the brightest pixel on the image set to better evaluate the eventual perfusion defects across the study. Analytically, this can be achieved by normalizing the brightest pixel of each image taking into account the brightest pixel image series as the Equation 3.2 and Equation 3.3 show.

**Table 3.4:** Train, validation and test data distribution.

| | | | n | % | total n | total % |
|---|---|---|---|---|---|---|
| Train | Normal | | 29 | 42.6 | 29 | 42.6 |
| | Abnormal | Ischemia | 29 | 42,6 | 39 | 57.4 |
| | | Infarct | 4 | 5.9 | | |
| | | Infarct and Ischemia | 6 | 8.9 | | |
| Validation | Normal | | 20 | 50 | 20 | 50 |
| | Abnormal | Ischemia | 17 | 42.5 | 20 | 50 |
| | | Infarct | 3 | 7.5 | | |
| Test | Normal | | 16 | 76.1 | 16 | 76.1 |
| | Abnormal | Ischemia | 4 | 19 | 5 | 23.8 |
| | | Infarct | 1 | 4.8 | | |

$$pixelMaxNorm = \frac{imgMax \times 255}{seriesMax} \tag{3.2}$$

$$imgNorm = \frac{img}{imgMax} \times pixelMaxNorm \tag{3.3}$$

where $seriesMax$ is the brightest pixel on the series of images of a certain cardiac axis, $imgMax$ is the brightest pixel on the image to normalize, $pixelMaxNorm$ is the brightest pixel on the future normalized image, $img$ is the image to normalize and $imgNorm$ is the image normalized.

Finally, we set a threshold of 100 to save only the images with their brightest pixel above this limit to exclude each plane's first and last slices because of the limited information they provide.

Due to the limited studies available to train the algorithm, and since DL models usually require large amounts of data, we proceeded to do data augmentation on the individual images we extracted. Thus, assuring a balanced training set at the end, we used rotations over 15°and crops to obtain individual images of shape 30x30.

With all the new augmented images, we followed to concatenate them into a single image. For this we created an image of shape 210x210, allowing in total of 49 individual images ((30x30)x7), and we added the individual images into this image, first the SA axis, then the HLA and finally the VLA. Due to time constraints and lack of availability, not all studies were

reconstructed post-Poisson subsampling. In this sense, the number of images of the datasets varied slightly from one training job to another. In Table 3.5 we can see the training data distribution for each label on each training job and the several testing sets used. For the testing and validation sets, after we extracted the images as previously stated, we followed to concatenate them into a single image as referred, without performing any data augmentation. The flow chart of how the data was created from the extraction of the cardiac slices in DICOM until the final input image is illustrated in Figure 3.1.

**Table 3.5:** Training set distribution for each training job and testing sets distribution across classes.

| | | labels | | | |
|---|---|---|---|---|---|
| | | Normal | | Abnormal | |
| | | n | % | n | % |
| training | 25S | 312 | 50.6 | 304 | 49.4 |
| | 50S | 319 | 50.6 | 312 | 49.4 |
| | 75S | 287 | 50.6 | 280 | 49.4 |
| | 100R | 319 | 50.6 | 312 | 49.4 |
| | ALL | 1237 | 50.6 | 1208 | 49.4 |
| testing | 25Rtest | 15 | 75 | 5 | 25 |
| | 25Stest | 15 | 79 | 4 | 21 |
| | 50Rtest | 15 | 75 | 5 | 25 |
| | 50Stest | 16 | 76.2 | 5 | 23.8 |
| | 75Rtest | 16 | 76.2 | 5 | 23.8 |
| | 75Stest | 16 | 76.2 | 5 | 23.8 |
| | 100Rtest | 15 | 75 | 5 | 25 |

Before feeding the data to the algorithm, in the training script, we used once again the data augmentation steps to further minimize the risk of overfitting. When pre-trained models with the ImageNET dataset are used, commonly the dataset is normalized to the mean and standard deviation of the data they were trained on. However, since our dataset is very distinct from the ImageNET one, we normalized our data based on the mean and standard deviation of the prepared dataset.

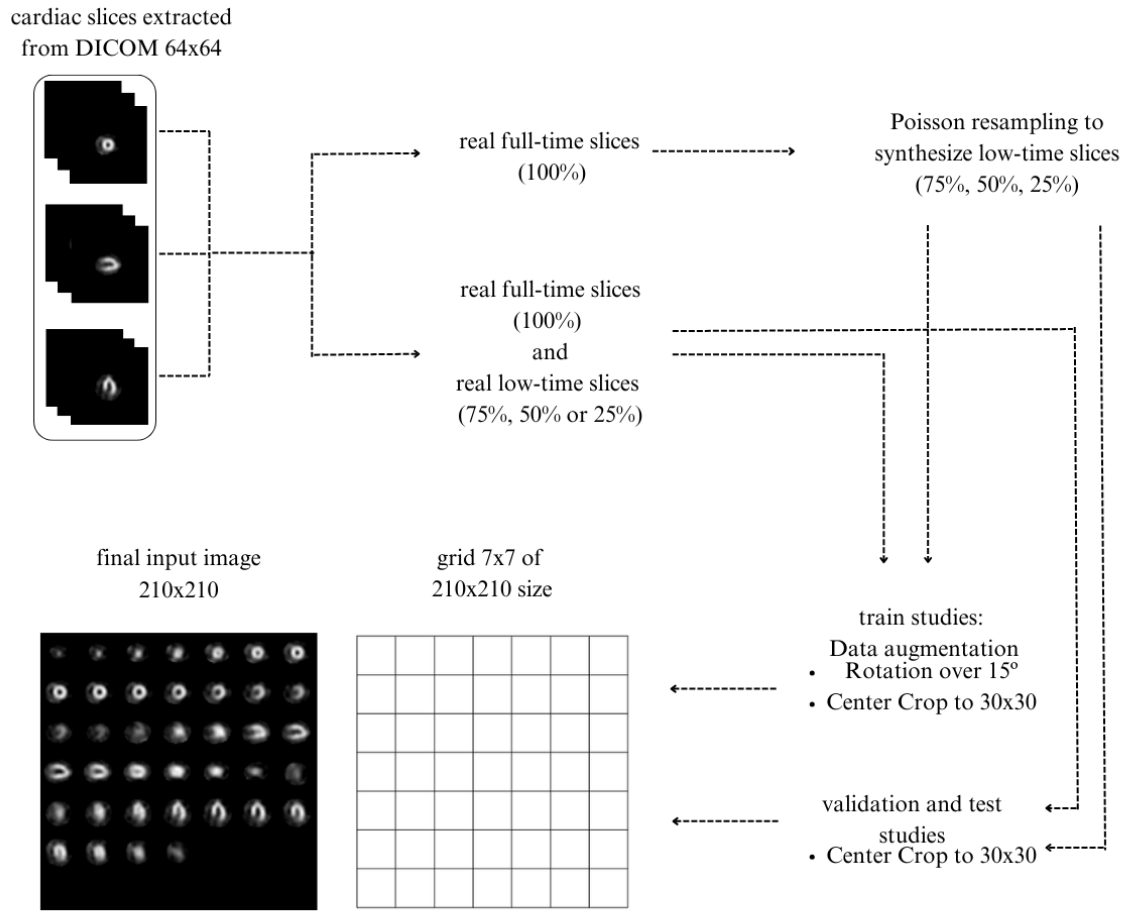



**Figure 3.1:** Flow chart of the data preparation.

### 3.5.3 Implementation Details

For these experiments, we trained the ResNet-18 model with pre-trained weights available in the Pytorch library. This model was initially trained in the ImageNET dataset to classify 1000 classes. In this sense, since our problem consists of binary classification (normal vs. abnormal), we modified the last fully connected layer of the model from 1000 to 2 output features (see Figure 3.2). Using the transfer learning technique, all the previous layers were frozen to not destroy any previously calculated weights during future training rounds.

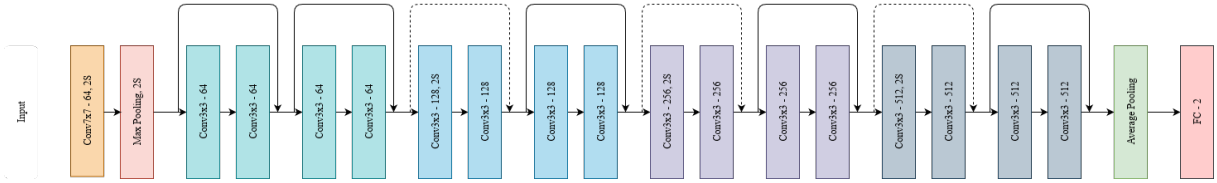



**Figure 3.2:** ResNet-18 architecture with 2 output classes.

The training jobs were run in the Google Colaboratory Research platform using Graphic Processing Units (GPU) T4. Data augmentation was implemented with *RandomRotation(degrees=15)*, for rotation, and *CenterCrop(30)*, for cropping purposes, from *pytorch.torchvision* library. All runs were set to train for 50 epochs using the Cross-entropy loss function, given by Equation 3.4. The batch size was 26 and 60 for all the individual sets of data and for all the previous ones combined, respectively. Regarding the learning rate, SGD and Adam were used, with a step size of 2 and a gamma of 0.1. These hyperparameters for each training job are described in Table 3.6.

$$\text{Cross-Entropy Loss} = -\frac{1}{N}\sum_{i=1}^{N}(t_i \cdot \log(p_i) + (1 - t_i) \cdot \log(1 - p_i)) \tag{3.4}$$

where $N$ is the number of samples in the dataset, $t_i$ is the true label for sample $i$, where $t_i = 1$ for the correct class and $t_i = 0$ for the other class, and $p_i$ is the predicted probability that sample $i$ belongs to the correct class.

**Table 3.6:** Training details of each training job

| | epochs | batch size | optimizer | learning rate | loss function |
|---|---|---|---|---|---|
| 25S | 50 | 26 | SGD | 0.01 (scheduler StepLR) | CrossEntropyLoss |
| 50S | 50 | 26 | SGD | 0.01 (scheduler StepLR) | CrossEntropyLoss |
| 75S | 50 | 26 | SGD | 0.01 (scheduler StepLR) | CrossEntropyLoss |
| 100R | 50 | 26 | Adam | 0.1 (scheduler StepLR) | CrossEntropyLoss |
| ALL | 50 | 60 | Adam | 0.001(scheduler StepLR) | CrossEntropyLoss |

### 3.5.4 Evaluation Metrics

To assess the performance of the classifiers, we measured both singular metrics and rank-based metrics. Singular metrics take as inputs the class predictions from the confusion matrix, without considering the actual classifier's output before being thresholded to the class prediction. Thus, although these metrics cannot provide an overall measure of the model's quality, they provide good insights into the model's performance. In this sense, we measured the accuracy,

sensitivity, specificity and f1-score. In Appendix B we present an extended explanation of the singular evaluation metrics used for this task.

Regarding rank-based metrics, these ones allow for an overall assessment of a model's performance, as they evaluate the model for different class decision threshold values. In this work, we calculated the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. Being ROC the relation between the true positive rate (TPR) and the false positive rate (FPR) and AUC the metric that shows how good a model is in discriminating between positive and negative target classes.

For explainability purposes of the model predictions, we also computed a Grad-CAM (Gradient-weighted Class Activation Mapping) analysis. In this approach, we can produce localization maps that give us hints of the important regions of the image the model used for class prediction. In this sense, Grad-CAMs are highly used in the DL field to improve its interpretability.

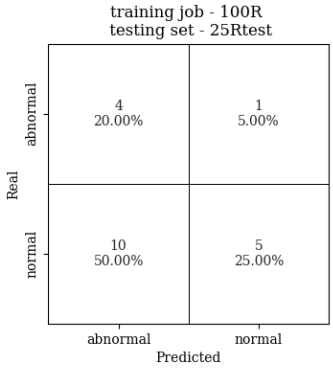
### 3.5.5 Results

Regarding the training job with the 100R, we can observe the resultant confusion matrices when testing with full-time count data (100Rtest), real and synthesized low-count data (75Rtest, 75Stest, 50Rtest, 50Stest, 25Rtest and 25Stest) in Figure 3.3. Additionally, the calculated singular metrics and AUC scores for every test set are represented in Table 3.7.

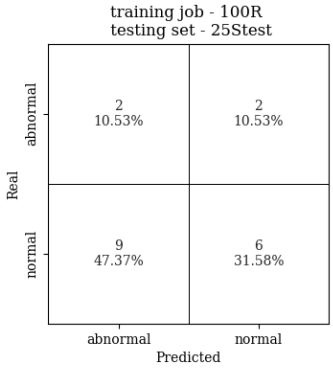
**Table 3.7:** Singular metrics and AUC scores for every test set in the 100R training job

| | Accuracy | Sensitivity | Specificity | F1-score | AUC-score |
|---|---|---|---|---|---|
| 25Rtest | 0.45 | 0.80 | 0.33 | 0.42 | 0.36 |
| 25Stest | 0.42 | 0.50 | 0.40 | 0.27 | 0.37 |
| 50Rtest | 0.65 | 0.80 | 0.60 | 0.53 | 0.73 |
| 50Stest | 0.38 | 0.40 | 0.38 | 0.24 | 0.30 |
| 75Rtest | 0.71 | 0.80 | 0.69 | 0.57 | 0.61 |
| 75Stest | 0.57 | 0.80 | 0.50 | 0.47 | 0.69 |
| 100Rtest | 0.70 | 0.80 | 0.67 | 0.57 | 0.73 |

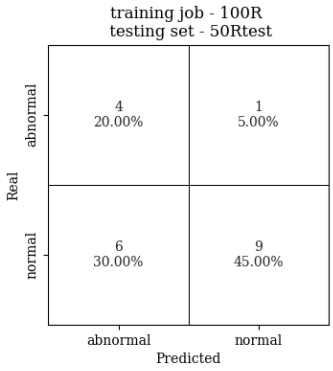
With respect to the individual training jobs 25S, 50S and 75S, the confusion matrices generated from the testing on the respective data count percentage are shown in Figure 3.4,



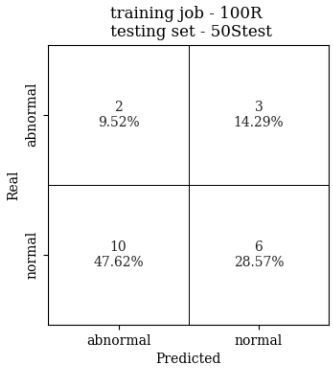

**(a)**



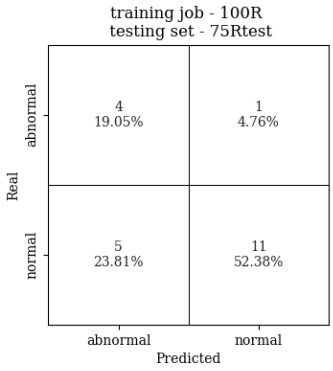

**(b)**



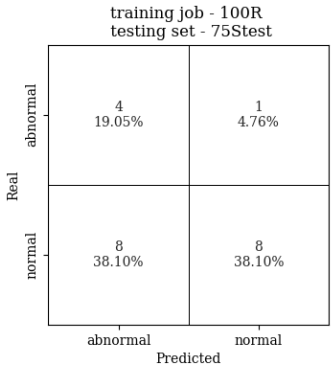

**(c)**



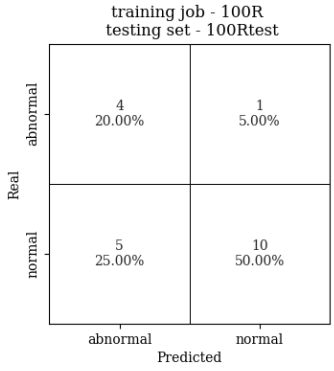

**(d)**




**(e)**




**(f)**




**(g)**

**Figure 3.3:** Confusion matrices generated from the training job 100R for the following testing sets: **(a)** 25Rtest, **(b)** 25Stest, **(c)** 50Rtest, **(d)** 50Stest, **(e)** 75Rtest, **(f)** 75Stest, **(g)** 100Rtest

Figure 3.5 and Figure 3.6, respectively. Moreover, the singular metrics and AUC score for 25S, 50S and 75S are represented in Table 3.8, Table 3.9 and Table 3.10, respectively.

**Table 3.8:** Singular metrics and AUC scores for the 25% test sets in the training job 25S

| | Accuracy | Sensitivity | Specificity | F1-score | AUC-score |
|---|---|---|---|---|---|
| 25Rtest | 0.65 | 0.80 | 0.60 | 0.53 | 0.68 |
| 25Stest | 0.58 | 0.25 | 0.67 | 0.20 | 0.40 |

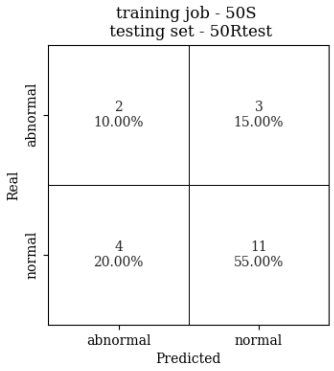
**Table 3.9:** Singular metrics and AUC scores for the 50% test sets in the training job 50S

| | Accuracy | Sensitivity | Specificity | F1-score | AUC-score |
|---|---|---|---|---|---|
| 50Rtest | 0.65 | 0.40 | 0.73 | 0.36 | 0.47 |
| 50Stest | 0.62 | 0.40 | 0.69 | 0.33 | 0.48 |

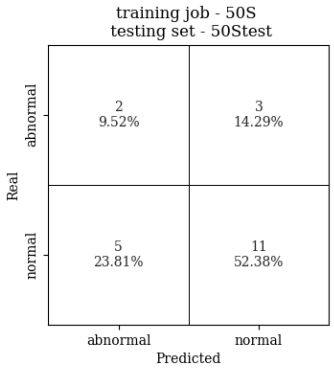
**Table 3.10:** Singular metrics and AUC scores for the 75% test sets in the training job 75S

| | Accuracy | Sensitivity | Specificity | F1-score | AUC-score |
|---|---|---|---|---|---|
| 75Rtest | 0.62 | 0.80 | 0.56 | 0.50 | 0.53 |
| 75Stest | 0.76 | 0.60 | 0.81 | 0.55 | 0.66 |


**Figure 3.4:** Confusion matrices generated from the training job 25S with the following testing sets: **(a)** 25Rtest, **(b)** 25Stest

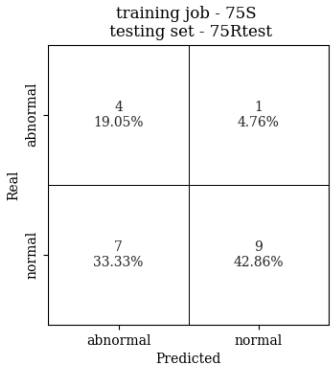



**(a)**

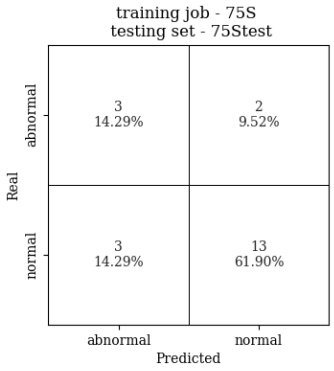



**(b)**

**Figure 3.5:** Confusion matrices generated from the training job 50S with the following testing sets: **(a)** 50Rtest, **(b)** 50Stest




**(a)**




**(b)**

**Figure 3.6:** Confusion matrices generated from the training job 75S with the following testing sets: **(a)** 75Rtest, **(b)** 75Stest

Finally, in regards to the training job ALL, the confusion matrices produced when tested with real and synthesized data across all percentages can be seen in Figure 3.7. In Table 3.11 we can find the evaluation scores when tested in all sets of percentages.

For explainability purposes of the performance of the several training jobs, we computed a Grad-CAM analysis for all the images in every testing set for all the training jobs. Given our goal to mainly compare the results between the 100R and ALL training jobs, Table 3.12 shows two images with an embedded Grad-CAM analysis, one from the "normal" class and another one from the "abnormal" class for all real testing sets.

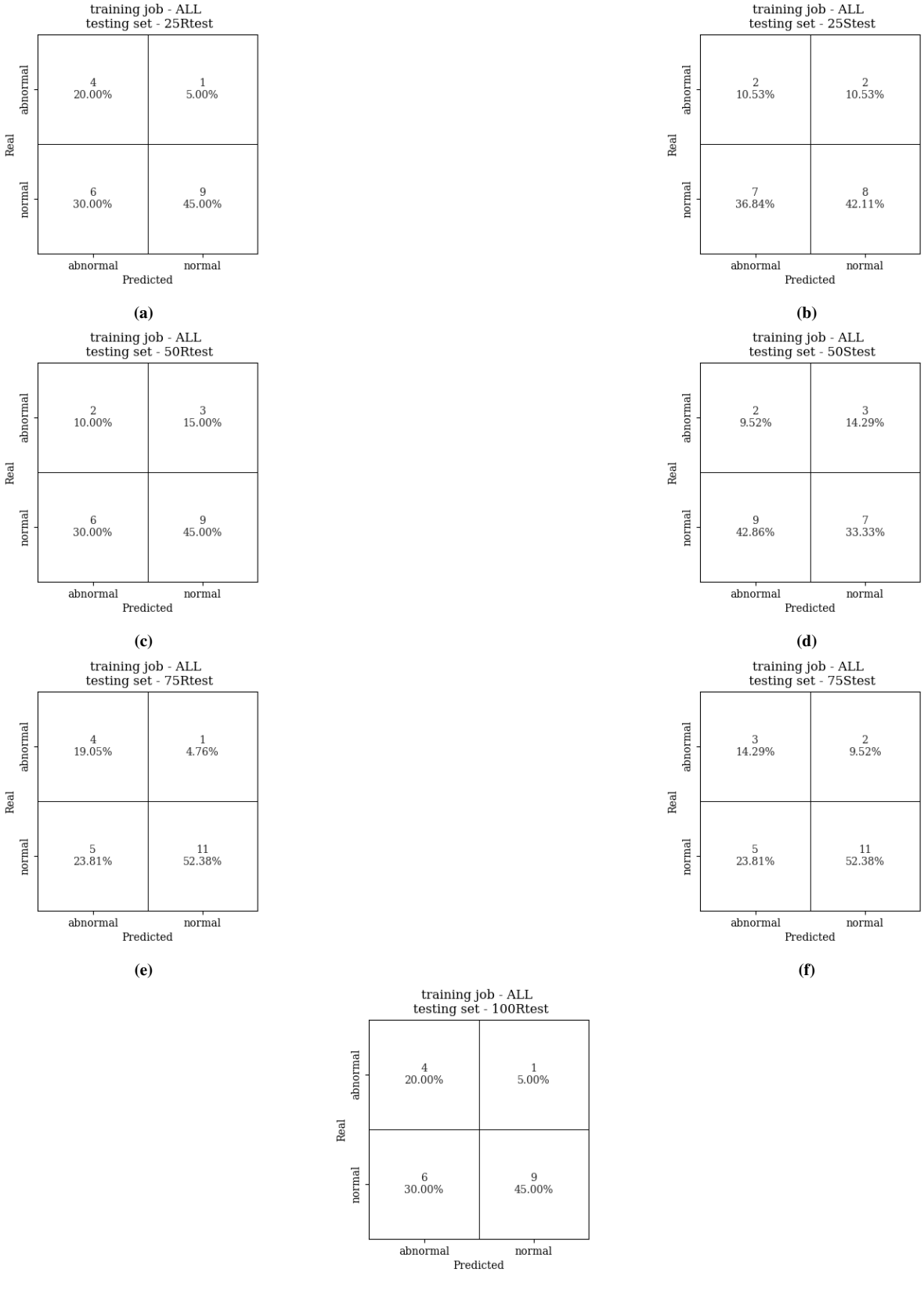



**Figure 3.7:** Confusion matrices generated from the training job ALL for the following testing sets: **(a)** 25Rtest, **(b)** 25Stest, **(c)** 50Rtest, **(d)** 50Stest, **(e)** 75Rtest, **(f)** 75Stest, **(g)** 100Rtest

**Table 3.11:** Singular metrics and AUC scores for every test set in the training job ALL

| | Accuracy | Sensitivity | Specificity | F1-score | AUC-score |
|---|---|---|---|---|---|
| 25Rtest | 0.65 | 0.80 | 0.60 | 0.53 | 0.69 |
| 25Stest | 0.53 | 0.50 | 0.53 | 0.31 | 0.52 |
| 50Rtest | 0.55 | 0.40 | 0.60 | 0.31 | 0.37 |
| 50Stest | 0.43 | 0.40 | 0.44 | 0.25 | 0.40 |
| 75Rtest | 0.71 | 0.80 | 0.69 | 0.57 | 0.68 |
| 75Stest | 0.67 | 0.60 | 0.69 | 0.46 | 0.69 |
| 100Rtest | 0.65 | 0.80 | 0.60 | 0.53 | 0.69 |

## 3.5.6 Discussion

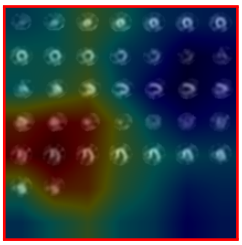
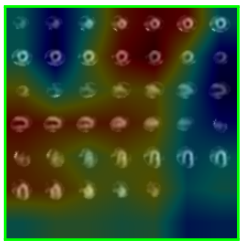
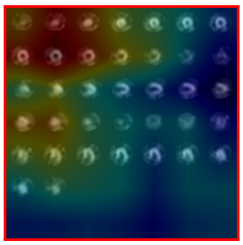
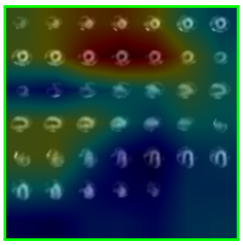
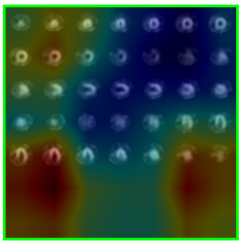
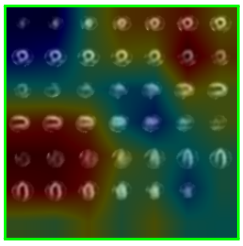
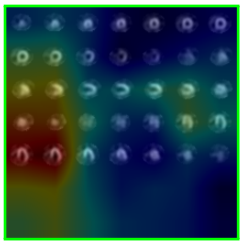
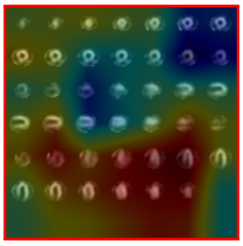
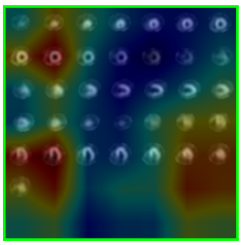
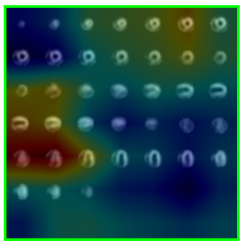
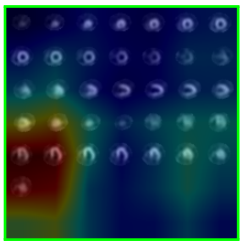
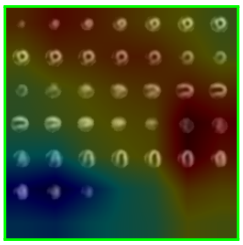
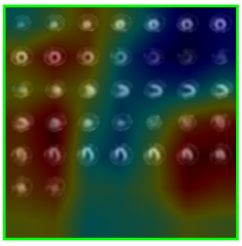
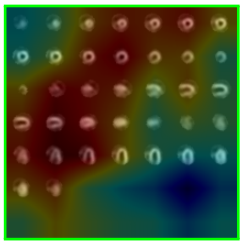
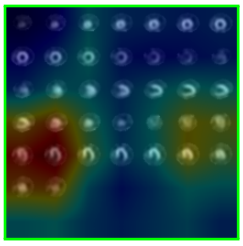
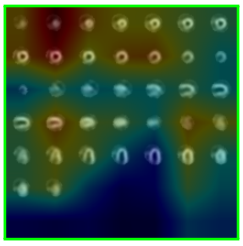
### 3.5.6.1 Poisson resampling results

In this work we generated synthetic images through the implementation of the Poisson resampling method described in [27], to further use in training and testing of the DL classifier. Despite the well-documented validation of the method in the literature [27, 22] either in static and tomographic acquisitions, we went to explore our results by calculating the boxplot to compare the pixel values distribution of real and subsampled data in the $30^{th}$ projection of a study and through the direct absolute difference of a real and subsampled middle-projection.

In the original paper on the method [27] they concluded the Poisson resampling method can accurately simulate the statistical nature of real static acquisitions, maintaining the correct mean, SD and skewness, over the full range of counts. Moreover, in [22] they synthesized low-count images with 67%, 50% and 33% count settings and investigated whether the dose of $^{201}$Tl during MPI could be reduced without deteriorating the quality of images. They concluded the data subsampled to 67% of the original data did not induce a significant loss of image quality, therefore, the dose of $^{201}$Tl could be reduced. However, quantification metrics such as SSS and SRS can be changed significantly in some cases.

In our brief analysis of the implementation of the Poisson resampling method, as can be seen in Figure A.2, the mean of the pixel values distribution is consistently slightly higher in the Poisson subsampled data in all percentages compared to the real raw-data acquisitions. These results may be possibly explained by to tracer's radioactive decay over time and the pharmacokinetics of the radiotracer since the synthetic images were generated from the full-

**Table 3.12:** Grad-CAM analysis for a normal and abnormal test image across all percentages for 100R and ALL training jobs. A green means the image was well predicted by the classifier. A red means the image was wrongly predicted by the classifier.

| | 100R | | ALL | |
|---|---|---|---|---|
| | True Label | | | |
| | Normal | Abnormal | Normal | Abnormal |
| 25Rtest |  |  |  |  |
| 50Rtest |  |  |  |  |
| 75Rtest |  |  |  |  |
| 100Rtest |  |  |  |  |

time studies. Accordingly, to understand where in the image these differences in pixel values are more prevalent, we computed the subtraction between three projections of a study. In Table A.1 we can see there are higher absolute difference values around the hepatobiliary system and the gastrointestinal tract, which goes in line with the previous statement, that is due to the pharmacokinetics of the tracer. This finding is a limitation of our implementation since we did not take into account the radioactive decay and the pharmacokinetics of the tracer over time. Future studies may include these corrections to better mimic the true count statistics of SPECT-MPI images.

#### 3.5.6.2 Results of the Deep Learning Classification SPECT-MPI

As intended for this project, we ran 5 training jobs to binary classify myocardial perfusion images standard full-time and synthesized low-count images between normal and abnormal (from ischemia or infarct).

Regarding our results in the classification task, the 100R (see Table 3.7) and ALL models (see Table 3.11) achieved accuracies of 0.70 and 0.65 and an AUC of 0.73 and 0.60 in the 100Rtest, respectively. Confronting the obtained results with the literature, in [18] they trained two different classification models, one DL-based and another expert knowledge-based to identify myocardial perfusion abnormalities using stress and rest images from 192 patients across training, validation and testing data. In the DL-based model, they used several pre-trained NNs being the ResNet-18 one of them, achieving an accuracy of 0.77, being slightly better than both our models. Moreover, in [92], they used two DL models to diagnose CAD. In this study, they used stress and rest SPECT-MPI images from 647 patients and classified them between "normal", "ischemic" and "infarction". To achieve this, they trained a CNN with 4 convolutional layers and for purposes of comparison, they also used two pre-trained models with transfer learning: VGG-16 and DenseNet-121. While the CNN model achieved an accuracy of 0.93 and an AUC of 0.94, the best pre-trained model was the VGG-16 with an accuracy of 0.89 and an AUC of 0.91. This study in particular achieved better accuracies and AUC compared to both our models. A possible explanation for this are: (1) they used different NN architectures, which can impact how well the NN fits the given data and (2) they also had a much larger dataset available containing images from 647 patients. Additionally, in [28], they developed a NN to detect CAD in SPECT-MPI at a threshold of 70% and 50% stenosis. Thus, an ANN was trained using 243 patients and tested in 65 patients achieving an accuracy of 0.72 and an AUC of 0.84 at 70% threshold and an

accuracy of 0.75 and an AUC of 0.84 at 50% threshold. Comparing these results to ours, their classifier achieved slightly better performance than both our models. However, their dataset contained also more images and they used a NN developed from scratch trained specifically for this classification problem. Contrarily, we used a model pre-trained on a different set of data and used it in a different classification problem, thus possibly explaining our sub-optimal results.

In regards to the reduced-count data, despite the majority of the studies referring low-dose MPI data for denoising purposes, some of them assess the diagnostic accuracy when reducing the administered activity. In [23], they used optimized reconstruction strategies on simulated reduced-count images (at 50%, 25% and 12.5%) and investigated whether reducing the administered activity could maintain the diagnostic accuracy in evaluating cardiac function. Thus, they concluded the activity could be reduced to 25% of the full dose without degrading the function assessment performance. However, to our knowledge, there are no studies regarding the classification of low-dose SPECT-MPI to further verify these findings and to assess the application of synthetic low-count data in classification tasks. Accordingly, in the three individual training jobs 25S, 50S and 75S, in Table 3.8, Table 3.9 and Table 3.10, respectively, we can see that the model trained with a synthesized set of images of a certain percentage can learn and classify accurately real and synthesized data of the same percentage, with some limitations. Thus, in the 75S training job, both tests achieved f1-scores of 0.50 and 0.55 in real and synthesized data, respectively. Regarding the 50S training job, the performance from both tests dropped the f1-scores to 0.36 and 0.33 and AUC scores of 0.47 and 0.48 on the 50Rtest and 50Stest, respectively. And, in the 25S model, the classifier achieved f1-scores of 0.53 and 0.20 in the 25Rtest and 25Stest sets, respectively.

Between the 100R model (see Table 3.7) and the ALL model (see Table 3.11), the overall results on the 75% tests were similar. Thus, both models achieved in the 75Rtest an accuracy of 0.71 and a f1-score of 0.57. In the 75Stest, the 100R model achieved an accuracy of 0.57, while the ALL model achieved 0.67. Concerning lower-count settings, in the 50Rtest the 100R model performed better, achieving an accuracy of 0.65 compared to 0.55 of the ALL model. In the 50Stest, the accuracy of both models dropped, achieving 0.38 and 0.43 for the 100R and ALL models, respectively. Regarding the 25Rtest and 25Stest, the ALL model performed better in both compared to the 100R model. Thus, achieving 0.65 and 0.53 in the 25Rtest and 25Stest, compared to 0.45 and 0.42, respectively, by the 100R model.

Given these results, and to better understand the decision-making of the classifiers, we

computed a Grad-CAM analysis for every training job. As an example, Table 3.12 illustrates a Grad-CAM analysis for two tested images (one normal and one abnormal) for the 100R and ALL training jobs. In this sense, although we can conclude that the models sometimes make the right prediction (shown by the green exterior line) taking into account important medial cardiac slices, for example in the 100Rtest in the 100R training job; at other times the models take too much into account slices with very limited information, for example, the abnormal image in the 100Rtest of the ALL training job. This fact may be due to the images' organization into the matrix grid in the data preparation process.

Despite these promising results in the classification of full-time and low-time data, there are several limitations of this study. Firstly, we had a very small dataset containing only 129 studies, using only 68 of them to training the models. However, it is known DL tasks require large amounts of data. To surpass this limitation, we used data augmentation. However, using a more diverse set of images, with more entry points, a more robust and generalizable final model is likely to be obtained. Secondly, our test sample was also very small and imbalanced, containing only 21 studies. This was mainly due to the extra time it takes to re-acquire patient images, which is not feasible in large scale, in addiction to the limited time we had to do this work. In this sense, it causes bias in the measured metrics towards the majority class. Thirdly, the difficulties of the classifiers to identify abnormalities can be possibly explained by the use of a sub-optimal filtering technique of the data. In this work, the data for the classification test was filtered by a wiener filter, however, this method has characteristics of both smoothing and blurring compensation [93]. However, the Butterworth filter is the recommended filtering technique by the International Atomic Energy Agency (IAEA) and may better balance the image quality by order and cut-off frequency parameterization [94].

# Chapter 4

# Conclusions and Future Work

The main goal of this thesis was to develop an effective image classification framework for standard full-time and low-time SPECT-MPI images while assessing the viability of feeding the DL models' by using the low-time data obtained from the Poisson resampling technique. To achieve this, we ran 5 training jobs: one with real full-time MPI images (100R) to assess its performance when tested with full-time and low-time data (real and synthetic), three individual training jobs with synthetic data of different percentages (75S, 50S and 25S) to assess the viability of using synthetic data as input and to understand their ability to classify real and synthetic data of the same percentage, and lastly, we ran a final model with all datasets combined to compare its results with the 100R model.

In this work, we concluded that the NN trained with full-time MPI (100R) achieves good results when tested with full-time data and real low-time data at 75%. However, below this percentage, the classifier's performance drops, possibly mainly due to the limited information these images have. Regarding the NN trained with real full-time and synthetic low-time data of all percentages (75%, 50%, 25%, as ALL), this one achieved similar performance results compared to the previous one stated when tested with full-time images and real low-time data at 75%, without showing any improvement in performance. Additionally, and similarly to the training job ran with full-time data, this classifier's performance dropped when tested with lower percentages (50% and 25%). Regarding the synthetic data experiments, we concluded that from the three individual training jobs and the ALL classifier compared to the 100R, a DL classifier trained with synthetic data can classify MPI images to certain extent with some limitations. Thus, we can affirm that the use of down-sampled data to mimic real low-time MPI images may be a possible solution to conduct further studies regarding the classification of low-time

SPECT-MPI images.

Despite the reasonable and promising results in the classification task, future work may include more training and test samples to achieve more robust and generalizable models. To achieve this, a possible solution may be the use of public SPECT-MPI datasets. Regarding the data, recommended filtering techniques, such as the Butterworth filtering, or even novel iterative reconstruction algorithms, such as OSEM, can be a possible way to optimize the quality of the input images. Finally, future work may use data-specific NN architectures to better fit into the specific myocardial perfusion imaging classification problem at hand.

# Bibliography


[1] Atherosclerosis;. Available from: https://www.hopkinsmedicine.org/health/conditions-and-diseases/atherosclerosis.

[2] Abbott BG, Case JA, Dorbala S, Einstein AJ, Galt JR, Pagnanelli R, et al. Contemporary Cardiac SPECT Imaging—Innovations and Best Practices: An Information Statement from the American Society of Nuclear Cardiology. Journal of Nuclear Cardiology. 2018;25(5):1847-60.

[3] Lecoq P. Detectors in Medicine and Biology. vol. 2; 2020.

[4] Jung H. Basic Physical Principles and Clinical Applications of Computed Tomography. Progress in Medical Physics. 2021;32(1):1-17.

[5] Fragoso P, Carl C, Univers O. Myocardial Perfusion Imaging - Image Processing and Software Myocardial Perfusion Imaging. 2016;(June).

[6] Kocayigit I, Can Y, Yilmaz S, Kilic H. Undeployed and uncrushed coronary stent in the proximal region of the normal left anterior descending artery revealed four years after intervention. Postepy w Kardiologii Interwencyjnej. 2015;11(3):256-7.

[7] Cedars Cardiac Processing Workstations;. Available from: https://www.ovissolutions.com/cedars-cardiac-processing-workstations/.

[8] Prasad MPS, Senthilrajan DA. Leaf Features Extraction for Plant Classification using CNN. International Journal of Advanced Research in Science, Communication and Technology. 2021;(March):148-54.

[9] Meißner P, Watschke H, Winter J, Vietor T. Artificial neural networks-based material parameter identification for numerical simulations of additively manufactured parts by material extrusion. Polymers. 2020;12(12):1-28.

[10] Papandrianos N, Papageorgiou E, Anagnostis A, Papageorgiou K. Bone metastasis classification using whole body images from prostate cancer patients based on convolutional neural networks application. PLoS ONE. 2020;15(8):1-28. Available from: http://dx.doi.org/10.1371/journal.pone.0237213.

[11] Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging. 2018.

[12] Yani M, Irawan B, Setiningsih C. Application of Transfer Learning Using Convolutional Neural Network Method for Early Detection of Terry's Nail. Journal of Physics: Conference Series. 2019;1201(1).

[13] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Proceedings of the IEEE International Conference on Computer Vision. 2017;2017-October:618-26.

[14] Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: A review. Medical Image Analysis. 2019;58.

[15] Vaduganathan M, Mensah GA, Turco JV, Fuster V, Roth GA. The Global Burden of Cardiovascular Diseases and Risk: A Compass for Future Health. Journal of the American College of Cardiology. 2022;80(25):2361-71.

[16] Sanchis-Gomar F, Perez-Quilis C, Leischik R, Lucia A. Epidemiology of coronary heart disease and acute coronary syndrome. Annals of Translational Medicine. 2016;4(13):1-12.

[17] Neumann FJ, Sechtem U, Banning AP, Bonaros N, Bueno H, Bugiardini R, et al. 2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromes. European Heart Journal. 2020;41(3):407-77.

[18] Kaplan Berkaya S, Ak Sivrikoz I, Gunal S. Classification models for SPECT myocardial perfusion imaging. Computers in Biology and Medicine. 2020;123(June):103893. Available from: https://doi.org/10.1016/j.compbiomed.2020.103893.

[19] Svensson A, Åkesson L, Edenbrandt L. Quantification of myocardial perfusion defects using three different software packages. European Journal of Nuclear Medicine and Molecular Imaging. 2004;31(2):229-32.

[20] Lecchi M, Malaspina S, Scabbio C, Gaudieri V, Del Sole A. Myocardial perfusion scintigraphy dosimetry: optimal use of SPECT and SPECT/CT technologies in stress-first imaging protocol. Clinical and Translational Imaging. 2016;4(6):491-8.

[21] Teles P, Trincão M, Alves F, Antunes V, Calado D, Cantinho G, et al. Evaluation of the Portuguese population exposure to ionizing radiation due to x-ray and nuclear medicine procedures from 2013 to 2017. Radiation Physics and Chemistry. 2020;172(February):108762. Available from: https://doi.org/10.1016/j.radphyschem.2020.108762.

[22] Kim IH, Lee SJ, An YS, Choi SY, Yoon JK. Simulating dose reduction for myocardial perfusion SPECT using a Poisson resampling method. Nuclear Medicine and Molecular Imaging. 2021;55(5):245-52. Available from: https://doi.org/10.1007/s13139-021-00710-w.

[23] Juan Ramon A, Yang Y, Wernick MN, Pretorius PH, Johnson KL, Slomka PJ, et al. Evaluation of the effect of reducing administered activity on assessment of function in cardiac gated SPECT. Journal of Nuclear Cardiology. 2020;27(2):562-72. Available from: https://doi.org/10.1007/s12350-018-01505-x.

[24] Acampa W, Buechel RR, Gimelli A. Low dose in nuclear cardiology: State of the art in the era of new cadmium-zinc-telluride cameras. European Heart Journal Cardiovascular Imaging. 2016;17(6):591-5.

[25] Song C, Yang Y, Wernick MN, Pretorius PH, King MA. Low-Dose Cardiac-gated SPECT studies using a Residual Convolutional Neural Network Department of Electrical and Computer Engineering , Illinois Institute of Technology , Chicago , IL Department of Radiology , University of Massachussetts Medical School , . 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). 2019;(Isbi):653-6.

[26] Ramon AJ, Yang Y, Pretorius PH, Johnson KL, King MA, Wernick MN. Improving Diagnostic Accuracy in Low-Dose SPECT Myocardial Perfusion Imaging with Convolutional Denoising Networks. IEEE Transactions on Medical Imaging. 2020;39(9):2893-903.

[27] White D, Lawson RS. A Poisson resampling method for simulating reduced counts in nuclear medicine images. Physics in Medicine and Biology. 2015;60(9):N167-76.

[28] Guner LA, Karabacak NI, Akdemir OU, Karagoz PS, Kocaman SA, Cengel A, et al. An open-source framework of neural networks for diagnosis of coronary artery disease from myocardial perfusion SPECT. Journal of Nuclear Cardiology. 2010;17(3):405-13.

[29] Lyon MC, Foster C, Ding X, Dorbala S, Spence D, Bhattacharya M, et al. Dose reduction in half-time myocardial perfusion SPECT-CT with multifocal collimation. Journal of Nuclear Cardiology. 2016;23(4):657-67.

[30] Queiroz CC, Machado MAD, Ximenes AAB, Pino AGS, Netto EM. Technical note: Partitioning of gated single photon emission computed tomography raw data for protocols optimization. Journal of Applied Clinical Medical Physics. 2022;23(3):1-7.

[31] World Heart Report 2023. World Heart Report 2023 Confronting the World ' S Number. 2023.

[32] Malakar AK, Choudhury D, Halder B, Paul P, Uddin A, Chakraborty S. A review on coronary artery disease, its risk factors, and therapeutics. Journal of Cellular Physiology. 2019;234(10):16812-23.

[33] Sirajuddin A, Mirmomen SM, Kligerman SJ, Groves D, Burke AP, Kureshi F, et al. Ischemic heart disease: Noninvasive imaging techniques and findings. Radiographics. 2021;41(4):990-1021.

[34] Timmis A, Townsend N, Gale CP, Torbica A, Lettino M, Petersen SE, et al. European society of cardiology: Cardiovascular disease statistics 2019. European Heart Journal. 2020;41(1):12-85.

[35] Hutton BF. The origins of SPECT and SPECT/CT. European Journal of Nuclear Medicine and Molecular Imaging. 2014;41(SUPPL. 1).

[36] Jaszczak RJ, Coleman RE. Single Photon Emission Computed Tomography (SPECT) Principles and Instrumentation. Progress in Clinical Radiology. 1985.

[37] Jain D, Aronow WS. Noninvasive Diagnostic Modalities for the Evaluation of Coronary Artery Disease. Elsevier Inc.; 2016. Available from: http://dx.doi.org/10.1016/B978-0-12-802385-3.00011-5.

[38] Zaret BL, Beller GA. Clinical Nuclear Cardiology; 2010.

[39] Wu J, Liu C. Recent advances in cardiac SPECT instrumentation and imaging methods. Physics in Medicine and Biology. 2019;64(6).

[40] Dorbala S, Ananthasubramaniam K, Armstrong IS, Chareonthaitawee P, DePuey EG, Einstein AJ, et al. Single Photon Emission Computed Tomography (SPECT) Myocardial Perfusion Imaging Guidelines: Instrumentation, Acquisition, Processing, and Interpretation. Journal of Nuclear Cardiology. 2018;25(5):1784-846. Available from: https://doi.org/10.1007/s12350-018-1283-y.

[41] Bruyant PP. Analytic and iterative reconstruction algorithms in SPECT. Journal of Nuclear Medicine. 2002;43(10):1343-58.

[42] Piekarski E, Manrique A, Rouzet F, Le Guludec D. Current Status of Myocardial Perfusion Imaging With New SPECT/CT Cameras. Seminars in Nuclear Medicine. 2020;50(3):219-26.

[43] Manabe O, Kikuchi T, Scholte AJHA, El Mahdiui M, Nishii R, Zhang MR, et al. Radiopharmaceutical tracers for cardiac imaging. Journal of Nuclear Cardiology. 2018;25(4):1204-36. Available from: https://doi.org/10.1007/s12350-017-1131-5.

[44] Henzlova MJ, Duvall WL, Einstein AJ, Travin MI, Verberne HJ. ASNC imaging guidelines for SPECT nuclear cardiology procedures: Stress, protocols, and tracers. Journal of Nuclear Cardiology. 2016;23(3):606-39.

[45] Verberne HJ, Acampa W, Anagnostopoulos C, Ballinger J, Bondt PD, Buechel RR, et al. EANM procedural guidelines for radionuclide myocardial perfusion imaging with SPECT and SPECT / CT; 2015.

[46] Van Djik JD. Image Quality and Radiation Dose in Cardiac Imaging; 2016.

[47] Desiderio MC, Lundbye JB, Baker WL, Farrell MB, Jerome SD, Heller GV. Current Status of Patient Radiation Exposure of Cardiac Positron Emission Tomography and Single-Photon Emission Computed Tomographic Myocardial Perfusion Imaging. Circulation Cardiovascular imaging. 2018;11(12):e007565.

[48] Cerqueira MD, Allman KC, Ficaro EP, Hansen CL, Nichols KJ, Thompson RC, et al. Recommendations for reducing radiation exposure in myocardial perfusion imaging. Journal of Nuclear Cardiology. 2010;17(4):709-18.

[49] Suzuki Y, Matsumoto N, Yoda S, Amano Y, Okumura Y. Optimal Protocol of Myocardial Perfusion Imaging for Reduction of Radiation Exposure. Annals of Nuclear Cardiology. 2022;8(1):109-12.

[50] da Silva LABN. A Reconstrução Tomográfica em SPECT com Utilização dos Novos Métodos Iterativos "WBR": Comparação com Métodos Analíticos; 2013.

[51] Hussani MTA, Ali MH, Hayani A. The Use of Filtered Back projection Algorithm for Reconstruction of tomographic Image The Use of Filtered Back projection Algorithm for Reconstruction of tomographic Image Abstract :. 2019;(January 2014).

[52] Lyra M, Ploussi A. Filtering in SPECT image reconstruction. International Journal of Biomedical Imaging. 2011;2011.

[53] Holly TA, Abbott BG, Al-Mallah M, Calnon DA, Cohen MC, Difilippo FP, et al. Single photon-emission computed tomography. Journal of Nuclear Cardiology. 2010;17(5):941-73.

[54] Vandenberghe S, D'Asseler Y, de Walle RV, Kauppinen T, Koole M, Bouwens L, et al. Iterative reconstruction algorithms in nuclear medicine. Computerized Medical Imaging and Graphics. 2001.

[55] Mahbod A. Comparison of FBP and Iterative Methods (SPECT DATA). 2015;(August). Available from: https://www.researchgate.net/publication/280918085.

[56] Slart RHJA, Williams MC, Juarez-Orozco LE, Rischpler C, Dweck MR, Glaudemans AWJM, et al. Position paper of the EACVI and EANM on artificial intelligence applications in multimodality cardiovascular imaging using SPECT/CT, PET/CT, and cardiac CT. European Journal of Nuclear Medicine and Molecular Imaging. 2021;48(5):1399-413.

[57] Uribe CF, Mathotaarachchi S, Gaudet V, Smith KC, Rosa-Neto P, Bénard F, et al. Machine learning in nuclear medicine: Part 1—Introduction. Journal of Nuclear Medicine. 2019;60(4):451-8.

[58] Bradley J Erickson, Korfiatis P, Akkus Z, Kline TL. Machine Learning for Medical Imaging. Radiographics. 2017.

[59] Zemouri R, Zerhouni N, Racoceanu D. Deep learning in the biomedical applications: Recent and future status. Applied Sciences (Switzerland). 2019;9(8).

[60] Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. vol. 8. Springer International Publishing; 2021. Available from: https://doi.org/10.1186/s40537-021-00444-8.

[61] Krittanawong C, Johnson KW, Rosenson RS, Wang Z, Aydar M, Baber U, et al. Deep learning for cardiovascularmedicine: A practical primer. European Heart Journal. 2019;40(25):2058-2069C.

[62] Litjens G, Ciompi F, Wolterink JM, de Vos BD, Leiner T, Teuwen J, et al. State-of-the-Art Deep Learning in Cardiovascular Image Analysis. JACC: Cardiovascular Imaging. 2019;12(8P1):1549-65.

[63] Huang SC, Pareek A, Jensen M, Lungren MP, Yeung S, Chaudhari AS. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. npj Digital Medicine. 2023;6(1).

[64] Lin Q, Man Z, Cao Y, Deng T, Han C, Cao C, et al. Classifying functional nuclear images with convolutional neural networks: A survey. IET Image Processing. 2020;14(14):3300-13.

[65] O'Shea K, Nash R. An Introduction to Convolutional Neural Networks. International Journal for Research in Applied Science and Engineering Technology. 2015;10(12):943-7.

[66] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016;2016-Decem:770-8.

[67] Sarwinda D, Paradisa RH, Bustamam A, Anggia P. Deep Learning in Image Classification using Residual Network (ResNet) Variants for Detection of Colorectal Cancer. Procedia Computer Science. 2021;179(2019):423-31. Available from: https://doi.org/10.1016/j.procs.2021.01.025.

[68] Kim HE, Cosa-Linan A, Santhanam N, Jannesari M, Maros ME, Ganslandt T. Transfer learning for medical image classification: a literature review. BMC Medical Imaging. 2022;22(1):1-13. Available from: https://doi.org/10.1186/s12880-022-00793-7.

[69] Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks.

[70] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2015;07-12-June-2015:1-9.

[71] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. 2015:1-14.

[72] Ibrahim R, Omair Shafiq M. Explainable Convolutional Neural Networks: A Taxonomy, Review, and Future Directions. ACM Computing Surveys. 2023;55(10).

[73] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning Deep Features for Discriminative Localization. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016;2016-December:2921-9.

[74] Skandarani Y, Jodoin PM, Lalande A. GANs for Medical Image Synthesis: An Empirical Study. Journal of Imaging. 2023;9(3):1-16.

[75] Mendes J, Pereira T, Silva F, Frade J, Morgado J, Freitas C, et al. Lung CT image synthesis using GANs. Expert Systems with Applications. 2023;215(August 2021):119350. Available from: https://doi.org/10.1016/j.eswa.2022.119350.

[76] Liu Y, Chen A, Shi H, Huang S, Zheng W, Liu Z, et al. CT synthesis from MRI using multi-cycle GAN for head-and-neck radiation therapy. Computerized Medical Imaging and Graphics. 2021;91(May):101953. Available from: https://doi.org/10.1016/j.compmedimag.2021.101953.

[77] Yu H, Yang LT, Zhang Q, Armstrong D, Deen MJ. Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives.

Neurocomputing. 2021;444:92-110. Available from: https://doi.org/10.1016/j.neucom.2020.04.157.

[78] Chrysostomou C, Koutsantonis L, Lemesios C, Papanicolas CN. Deep Convolutional Neural Network for Low Projection SPECT Imaging Reconstruction. 2020 IEEE Nuclear Science Symposium and Medical Imaging Conference, NSS/MIC 2020. 2020:2-5.

[79] Turečková A, Tureček T, Komínková Oplatková Z, Rodríguez-Sánchez A. Improving CT Image Tumor Segmentation Through Deep Supervision and Attentional Gates. Frontiers in Robotics and AI. 2020;7(August):1-14.

[80] Sendra-Balcells C, Salvador R, Pedro JB, Biagi MC, Aubinet C, Manor B, et al. Convolutional neural network MRI segmentation for fast and robust optimization of transcranial electrical current stimulation of the human brain. bioRxiv. 2020:2020.01.29.924985. Available from: https://www.biorxiv.org/content/10.1101/2020.01.29.924985v1.full.

[81] Sohlberg A, Kangasmaa T, Constable C, Tikkakoski A. Comparison of deep learning-based denoising methods in cardiac SPECT. EJNMMI Physics. 2023;10(1). Available from: https://doi.org/10.1186/s40658-023-00531-0.

[82] Guo Y, Xu J, Li X, Zheng L, Pan W, Qiu M, et al. Classification and Diagnosis of Residual Thyroid Tissue in SPECT Images Based on Fine-Tuning Deep Convolutional Neural Network. Frontiers in Oncology. 2021;11(October):1-13.

[83] Lin Q, Man Z, Cao Y, Wang H. Automated Classification of Whole-Body SPECT Bone Scan Images with VGG-Based Deep Networks. International Arab Journal of Information Technology. 2023;20(1):1-8.

[84] Wang J, Fan X, Qin SS, Shi K, Zhang H, Yu F. Exploration of the efficacy of radiomics applied to left ventricular tomograms obtained from D-SPECT MPI for the auxiliary diagnosis of myocardial ischemia in CAD. International Journal of Cardiovascular Imaging. 2022;38(2):465-72. Available from: https://doi.org/10.1007/s10554-021-02413-x.

[85] Arsanjani R, Xu Y, Dey D, Vahistha V, Shalev A, Nakanishi R, et al. Improved accuracy of myocardial perfusion SPECT for detection of coronary artery disease by machine learning in a large population. Journal of Nuclear Cardiology. 2013;20(4):553-62.

[86] Lindahl D, Toft J, Hesse B, Palmer J, Ali S, Lundin A, et al. Scandinavian test of artificial neural network for classification of myocardial perfusion images. Clinical Physiology. 2000;20(4):253-61.

[87] Lindahl D, Palmer J, Ohlsson M, Peterson C, Lundin A, Edenbrandt L. Automated interpretation of myocardial SPECT perfusion images using artificial neural networks. Journal of Nuclear Medicine. 1997;38(12):1870-5.

[88] Cho J, Lee K, Shin E, Choy G, Do S. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? 2016. Available from: http://arxiv.org/abs/1511.06348.

[89] Hsiao EM, Cao X, Zurakowski D, Zukotynski KA, Drubach LA, Grant FD, et al. Reduction in radiation dose in mercaptoacetyltriglycerine renography with enhanced planar processing. Radiology. 2011;261(3):907-15.

[90] Aghakhan Olia N, Kamali-Asl A, Hariri Tabrizi S, Geramifar P, Sheikhzadeh P, Farzanefar S, et al. Deep learning–based denoising of low-dose SPECT myocardial perfusion images: quantitative assessment and clinical performance. European Journal of Nuclear Medicine and Molecular Imaging. 2022;49(5):1508-22. Available from: https://doi.org/10.1007/s00259-021-05614-7.

[91] Markiewicz PJ, Reader AJ, Matthews JC. Assessment of bootstrap resampling performance for PET data. Physics in Medicine and Biology. 2015;60(1):279-99. Available from: http://dx.doi.org/10.1088/0031-9155/60/1/279.

[92] Papandrianos NI, Feleki A, Papageorgiou EI, Martini C. Deep Learning-Based Automated Diagnosis for Coronary Artery Disease Using SPECT-MPI Images. Journal of Clinical Medicine. 2022;11(13).

[93] Masoomi MA, Al-Shammeri I, Kalafallah K, Elrahman HMA, Ragab O, Ahmed E, et al. Wiener filter improves diagnostic accuracy of CAD SPECT images—comparison to angiography and CT angiography. Medicine (United States). 2019;98(4).

[94] Sayed IS, Ismail SS. Comparison of Low-Pass Filters for SPECT Imaging. International Journal of Biomedical Imaging. 2020;2020.

[95] Vakili M, Ghamsari M, Rezaei M. Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification. 2020. Available from: http://arxiv.org/abs/2001.09636.

# Appendix A

# The Poisson Resampling Method - supplementary material

## A.1 Poisson subsampling function

In Figure A.1 is visible the function used to subsample the standard full-time SPECT-MPI studies with 25%, 50% and 75% count settings, as described in [27].

```
def poisson_resampling(true_data, prob):

    data_new = true_data.copy()

    prob = prob

    for proj in tqdm(range(60)):
        for x in range(64):
            for y in range(64):
                COUNTS = data[proj][x,y]
                if COUNTS == 0:
                    data_new[proj][x,y] = 0
                else:
                    for COUNT in range(COUNTS):
                        u = np.random.uniform(0,1)
                        if u <= prob:
                            pass
                        else:
                            COUNTS = COUNTS - 1
                    data_new[proj][x,y] = COUNTS
    return data_new
```

**Figure A.1:** Poisson resampling function used to subsampled MPI studies.

Where *true_data* is the full-time study from which we synthesize the low-time images, and *prob* is the percentage of the low-time data to be generated (0.25 for 25%, 0.5 for 50% and 0.75

to 75%).

### A.1.1 Validation

In the original paper [27] introducing the Poisson resampling technique for simulating reduced counts in nuclear medicine images, half-count images were generated from the original raw data. Thus, the mean, standard deviation and skewness of the distribution of all pixels on both samples were calculated within a standardised region of interest containing the useful field of view and excluding pixels near the edge of the field of view. In the end, they concluded the Poisson resampling method accurately simulates the statistical nature of a real acquisition, maintaining the correct mean, standard deviation and skewness, of the full range of counts.

Additionally, to better evaluate our Poisson resampling implementation, we calculated the boxplot for the $30^{th}$ projection for 25%, 50% and 75%, real and subsampled to further compare the data distribution between the real acquisitions and the subsampled ones. In Figure A.2 we can observe that the means of all distributions are very similar when compared to the real and Poisson subsampled studies of each percentage. Despite these similarities, the subsampled data presents higher values than the real acquisition.

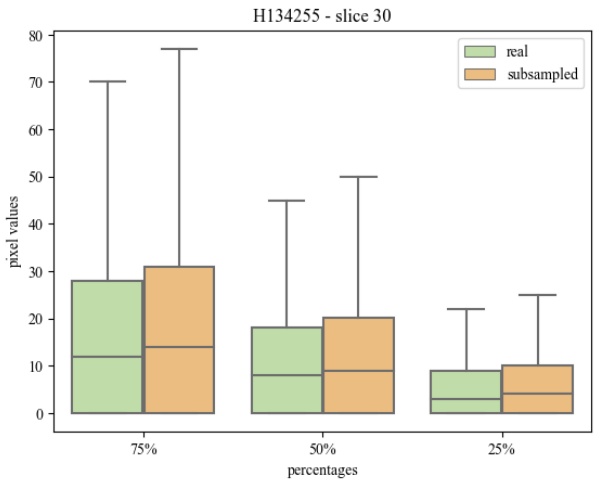
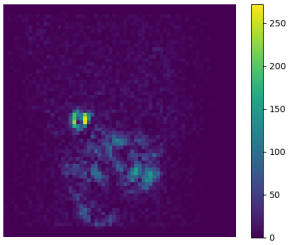
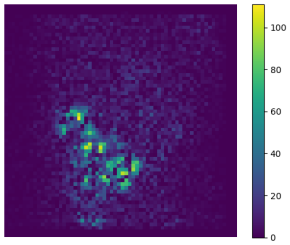
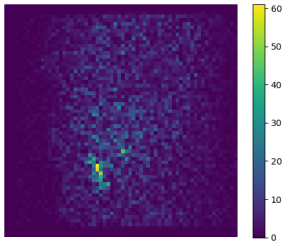
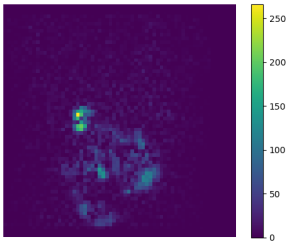
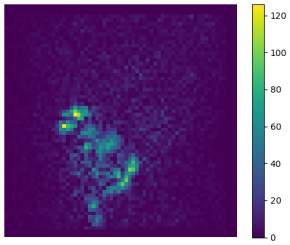
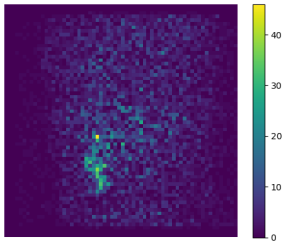
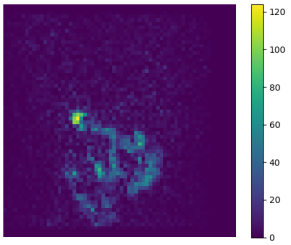
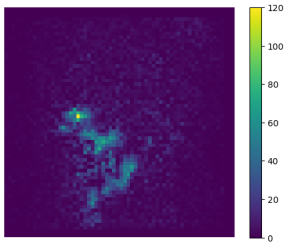
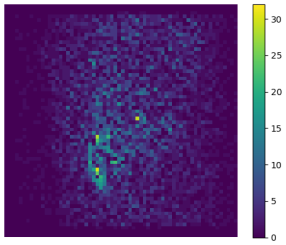



**Figure A.2:** Boxplot without outliers of the total counts of each percentage both real and subsampled

Given these results, and to understand these slight differences in data distributions, we computed the absolute difference between three projections (a RAO, a middle and a LPO projection) and the subsampled one of a study. In Table A.1, we can conclude that the pixels with the highest absolute difference correspond to the small bowel. Regarding the myocardial area, no great differences are visible.

**Table A.1:** Difference between real and subsample projections at different angles

| | RAO | middle | LPO |
|---|---|---|---|
| 75% |  |  |  |
| 50% |  |  |  |
| 25% |  |  |  |

# Appendix B

# Singular Evaluation Metrics Analitically

As said in subsection 3.5.4, in this work we used several singular evaluation metrics to evaluate the model's performance. Before explaining each of them and showing them analytically, it is helpful to introduce first the concept of a confusion matrix, which summarizes the classifier's prediction performance. Figure B.1 shows the matrix structure with its labels, where the rows represent the samples' true class, and the columns represent the classifier's class predictions.

If a sample belongs to the positive (P) class and the classifier correctly predicts the positive class, it corresponds to a true positive (TP). Otherwise, if the classifier predicts the negative class, it corresponds to a false negative (FN). If a sample belongs to the negative (N) class and the classifier correctly predicts its class, it corresponds to a true negative (TN). Otherwise, an incorrect prediction corresponds to a false positive (FP).

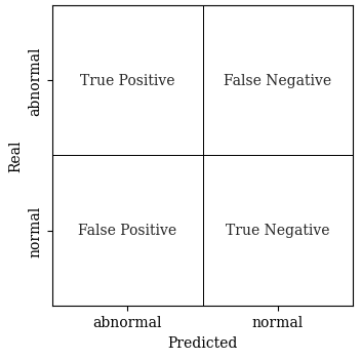



**Figure B.1:** Confusion matrix for a binary classification.

In this sense, singular metrics take as inputs the class predictions from the confusion matrix, without considering the actual classifier's output before being thresholded to the class prediction. Thus, although these metrics cannot provide an overall measure of the model's quality, they

provide good insights into the model's performance. In this work, we measured the accuracy, sensitivity, specificity and f1-score.

- Accuracy: It's the most known and used classification metric for evaluating algorithm performance in classification problems. It's defined as the ratio of correct classified items to the total number of observations, as shown in Equation B.1. It is important to note, that this metric can be an unreliable measure of model performance in imbalanced datasets due to its value can be near the maximum at 1.0, showing a large bias towards the negative class [95].

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{B.1}$$

- Sensitivity: Regarding the minority class, the sensitivity assesses out of the observations that are actually positive, how many of them have been correctly predicted by the algorithm, thus seeking to maximize the number of true positives, penalizing the false negatives. This metric is also commonly referred to as recall and it's calculated as Equation B.2 shows.

$$Sensitivity = \frac{TP}{TP + FN} \tag{B.2}$$

- Specificity: Contrary to the sensitivity metric, the specificity takes into account the majority class, showing the ability to detect a true negative. Analytically it is defined by Equation B.3.

$$Specificity = \frac{TN}{TN + FP} \tag{B.3}$$

- F1-score: The f1-score is defined by the harmonic mean of the precision and recall scores. It's commonly used specifically in imbalanced datasets due to its clear bias towards the positive class and it is calculated as Equation B.4 shows.

$$F1 - score = \frac{2 \times precision \times sensitivity}{TP + FN + TN + FP} \tag{B.4}$$