UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S THESIS

Hand-crafted and Deep Learning-based CMR Radiomics for Diagnosis of Cardiovascular Diseases

Author:
Alejandro HERNANDEZ

Supervisor
Dr. Polyxeni GKONTRA
Akshay JAGGI
Dr. Karim LEKADIR

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamental Principles of Data Science

in the

Facultat de Matemàtiques i Informàtica

September 1, 2020

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

Hand-crafted and Deep Learning-based CMR Radiomics for Diagnosis of Cardiovascular Diseases

by Alejandro HERNANDEZ

Cardiovascular diseases (CVDs) are subject of interest among researchers and clinicians due its high mortality rate. Cardiac Magnetic resonance (CMR) is the reference for clinicians to analyze the heart tissues by visual assessment and crude quantitative measures of the structures to dictate a diagnosis of the patient's status. Radiomics is a novel image analysis technique extract a large number of quantitative features from CMR that provide insightful information of the heart structures to support clinicians in the diagnosis and prognosis of these diseases. Previous studies have demonstrated the capacity of these features to obtain higher diagnosis accuracy than conventional methods. In this study, we cover in-depth Radiomics technique and explore two types of radiomics: Hand-Crafted Radiomics (HCR) and Deep Learning-based Radiomics (DLR). HCR computes a wide range of researcher-defined quantitative features that measure the shape, intensity, and texture of image regions of interest. . DLR are features extracted based in the training of Convolutional Neural Networks (CNNs). We address the methodology for the extraction of the features for both methods and analyze the performance in the CVDs classification task with Machine Learning (ML) algorithm. We also develop a pipeline for the fusion of these features with the aim of collecting complementary information of the heart structures from both mentioned methods with the aim of improving diagnostic accuracy. We apply this methodology with two benchmark medical datasets: ACDC Challenge dataset and UK BioBank with the availability of both CMR and the segmentation of the heart structures: Myocardium (MYO), Right Ventricle (RV) and Left Ventricle (LV). We perform an analysis of the results, discuss challenges and elaborate on future work.

Acknowledgements

I would like to thank my family and friends for their support never-ending support. I also want to thank all my supervisors: Xenia, Akshay, and Karim. First, for giving me the opportunity to work in such an important and interesting topic as it is the biomedical field and apply the concepts learned during the master as a tool to solve actual questions. And second, for devoting a great deal of time in my guidance during the realization of this study. Special appreciation to my direct supervisor Xenia, who always gave me very insightful feedback on my work, technical support, and lots of her time to help me during the study. I am extremely thankful for all her help. Also, I want to thank Alejandro Gonzalez, with I worked at the first stage of this project, in which we discussed and explored many methods that ended up being part of the project. I also want to thank Victor, Carlos, Christian and Katherina which, together with my supervisors, are part of the Laboratory of Artificial Intelligence in Medicine Lab of the UB. They provided us with data, segmentation, knowledge, and technical support when needed. Finally, I want to thank the master's professors and tutors for their efforts and adaptation during this complicated time.

Chapter 1

Introduction

1.1 Introduction

Cardiovascular diseases (CVD) are a topic of great interest in the medical research field because they remain one of the highest causes of mortality, accounting for approximately one-third of annual deaths (Martin-Isla et al., 2020). Early and precise diagnosis of these diseases can lead to more suitable and effective treatment, increasing the patient survival possibilities. In this context, cardiovascular magnetic resonance (CMR) imaging plays an important role in the diagnosis. Current CMR analysis techniques consist mainly of visual assessment, or, at best, in the extraction of few quantitative metrics of cardiac structure ad function, such as ejection fraction, end-systolic (ES) and end-diastolic (ED) volumes of the ventricles and myocardium. These classic techniques are tedious, time-consuming, prone to subjectivity, while they exploit only part of the information present in CMR. In recent years, with the high development of computational power, there has been an important presence of modern data analysis techniques and Artificial Intelligence (AI) in the medical field. Medical images have been an important focus for many researchers and data scientists. Machine learning (ML), a subfield of AI, has been used for image-based diagnosis by means of statistical algorithms that learn from past observations through the identification of hidden and complex imaging patterns.(Afshar et al., 2018)

Radiomics is a technique that involves calculating a large number of imaging descriptors from delineated images. Radiomics have been widely used in cancer imaging 1.1. In the context of CMR, radiomics describing changes in image appearance due to CVDs are attracting increasing interest. The large number of radiomics extracted consist of image quantification of shape and tissue character (Raisi-Estabragh et al., 2020). These features have been used to assess the predictive power for CVD and more. These features are known as hand-crafted radiomics (HCR). Lately, every year, radiomics research publications have been increasing and it is expected that its usage becomes an important complement for medical treatments in the future. (Afshar et al., 2018) Figure 1.1

Deep learning-based radiomics (DLR) consist of using deep learning models to automatically learn and extract meaningful features for performing the classification task at hand without having to explicitly tell the model which features to use. Deep Learning (DL) models, usually trained with large amounts of data, are capable to map the non-linearity with the target label. The mapping yields a highly sophisticated feature representation for the images, which is the key advantage of Convolutional Neural Networks (CNN), compared to other machine learning (ML) methods (Martin-Isla et al., 2020). These learned features can show better performance than

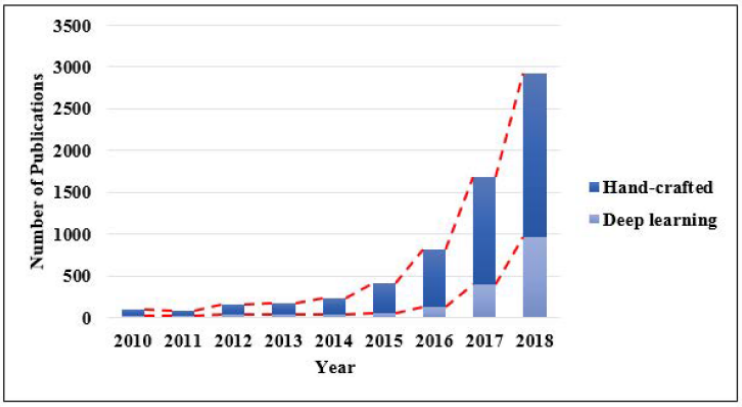

FIGURE 1.1: Radiomics publications in the latests years (Afshar et al., 2018)

HCR. Nevertheless, deeply learned classification does not provide enough explainability for decision-makers to understand the logic behind the classification (Moreira et al., 2020).

The innovation of this thesis mainly consists of performing an in-depth study of the performance of HCR and DLR for the classification of a wide range CVDs, such as hypertension, atherosclerotic, dilated cardiomyopathy, among others. In addition, develop a pipeline for fusing these two types of radiomics features to analyze and evaluate the performance of each method. In sum, the main aims of this thesis are:

- Automatically detect clinically relevant features from cardiovascular magnetic resonance for improved disease diagnosis by leveraging state-of-the art deep learning approaches.
- Explore the potential of deep learning-based radiomics as novel imaging biomarkers.
- Evaluate the integration of hand-crafted, deep learning-based radiomics and clinical information to improve automatic disease diagnosis.

We implement our methodology with two different medical datasets: (1) ACDC dataset, (2) UK Bio Bank. The ACDC dataset was created for the Automated Cardiac Diagnosis Challenge with data recorded from real clinical exams at the University Hospital of Dijon (France) for which there is already benchmark results (Bernard et al., 2018). The UK Bio Bank dataset was used with the intention of evaluating the applied methodology with a larger dataset and multiple diseases to ensure generalization of the performance of the methodology. UK Biobank is an international resource that contains data from half-million participants, including CMR for a large number of them. This database is highly used by the research community in the field of biomedicine and it is currently used for the European project *euCanSHare* [1].

It is certain that as the years pass, the involvement of AI techniques in the medical field will be larger and will benefit both clinicians and patients. This thesis is intended to mark an important step in this direction.

[1] http://www.eucanshare.eu/

1.2 Related Work

AI in Cardiovascular Imaging

Existing work demonstrates the incremental value of image-based cardiovascular diagnosis with ML for several important conditions such as coronary artery disease (CAD) and heart failure (HF). In these publications, researchers demonstrate the benefits of ML and DL for medical tasks. More precisely, (Cetin et al., 2017), extracted HCR for the ACDC Challenge dataset and applied Support Vector Machines (SVMs) for the classification of CVDs using sequential forward feature selection for identifying the most important HCR features. (Yang et al., 2018) used deep learning methods in combination with three functional measures (the amplitude of low-frequency fluctuations, regional homogeneity and regional functional correlation strength) based on rs-fMRI data to distinguish between migraineurs and healthy controls. They compare the performance of AlexNet modules and InceptionV3 as CNN architectures. Also, (Wang et al., 2017) compared the accuracy classification of mediastinal lymph node metastasis of non-small cell lung cancer of several machine learning algorithm as SVMs, Random Forests (RFs), Adaptive Boosting (AB) and CNNs, with human doctor's accuracy. For the ML classical methods they extracted texture features (HCR) from PET/CT scan and diagnostic features such as tumor size, CT value, SUV, image contrast, and intensity standard deviation.

Radiomics and Deep Learning Features Fusion

More recently, the first studies using a combination of HCR and DLR have started to appear in literature. (Lai and Deng, 2018) combine deep learning features with traditional statistical features from the medical images using a multilayer perceptron model. Their methodology is tested with two benchmark medical datasets: HIS2828 and ISIC2017. (Nie et al., 2019) developed a multimodal learning framework combining 3D deep learning features with traditional hand-crafted features to predict overall survival time of brain tumor patients. The authors obtained favorable results and improved survival time prediction. (Bizzego et al., 2019) developed a pipeline for the combination and unification of DLR and HCR features for the classification of patients with cancer, obtaining better results over published papers with same data.

Table 1.1 provides a list of relevant publications regarding the use of HCR and DLR along with the attached accuracy.

TABLE 1.1: Relevant Work and Performances

Publication	Method	HCR	DLR	Fusion
(Nie et al., 2019)	3D CNN + SVM	77.5	87	90.66
(Bizzego et al., 2019)	3D CNN + SVM	88	79.9	96.5
(Lai and Deng, 2018)	MLP combination of HCR+DLR	72.2	79.5	90.2
		66.1	75	90.1
(Wang et al., 2017)	Texture Features (HCR) and CNN	80-85	86	N/A
(Cetin et al., 2017)	HCR	0.92	N/A	N/A
(Yang et al., 2018)	DLR	N/A	86-99	N/A

Chapter 2

Datasets

In this section, we elaborate on medical databases to which we had access and with which we could work our methodology. We also provide a short description of the diseases we have worked with.

2.1 ACDC Challenge Dataset

The ACDC dataset was used for the Automated Cardiac Diagnosis Challenge (ACDC) Challenge [1]. The dataset is originated by real clinical exams at the university Hospital of Dijon (France). The dataset follows well defined pathologies in cine-MRI and the data is found in Nifti format. The dataset contains information from 150 patients, divided evenly in five groups, four diseases and one group of normal patients. The classes were equally distributed in the *Train* and *Test* datasets provided by the challenge organizer.

- 30 patients with dilated cardiomyopathy (diastolic left ventricular volume >100 mL/m2 and an ejection fraction of the left ventricle lower than 40%) - (**DCM**).
- 30 patients with previous myocardial infarction (ejection fraction of the left ventricle lower than 40% and several myocardial segments with abnormal contraction) - (**MINF**).
- 30 patients with hypertrophic cardiomyopathy (left ventricular cardiac mass high than 110 g/m2, several myocardial segments with a thickness higher than 15 mm in diastole and a normal ejection fraction) - (**HCM**).
- 30 patients with abnormal right ventricle (volume of the right ventricular cavity higher than 110 mL/m2 or ejection fraction of the right ventricle lower than 40%) – (**RV**).

Each group was clearly defined according to physiological parameter, such as the left or right diastolic volume or ejection fraction, the local contraction of the left ventricle, the left ventricle mass and the maximum thickness of the myocardium. Besides the cine-MRI, medical information such as height and weight were also given as information of the patients. Additionally, the segmentation of the region of interest as are the Myocardium (MYO), Right Ventricle (RV), and left ventricle (LV) are available. In table 2.1 is shown the distribution of the groups for train and test dataset. In Figure 2.1 we show in the first row, from left to right, the original Cine-MRI, the manually segmentation of the heart structures and the combination of the Cine-MRI and Mask. In the second row we show RV, MYO and LV.

[1] ACDC Challenge : https://acdc.creatis.insa-lyon.fr/description/

TABLE 2.1: ACDC Dataset.

Dataset	No. Patients	DCM	HCM	MINF	RV	NOR
Train	100	20	20	20	20	20
Test	50	10	10	10	10	10

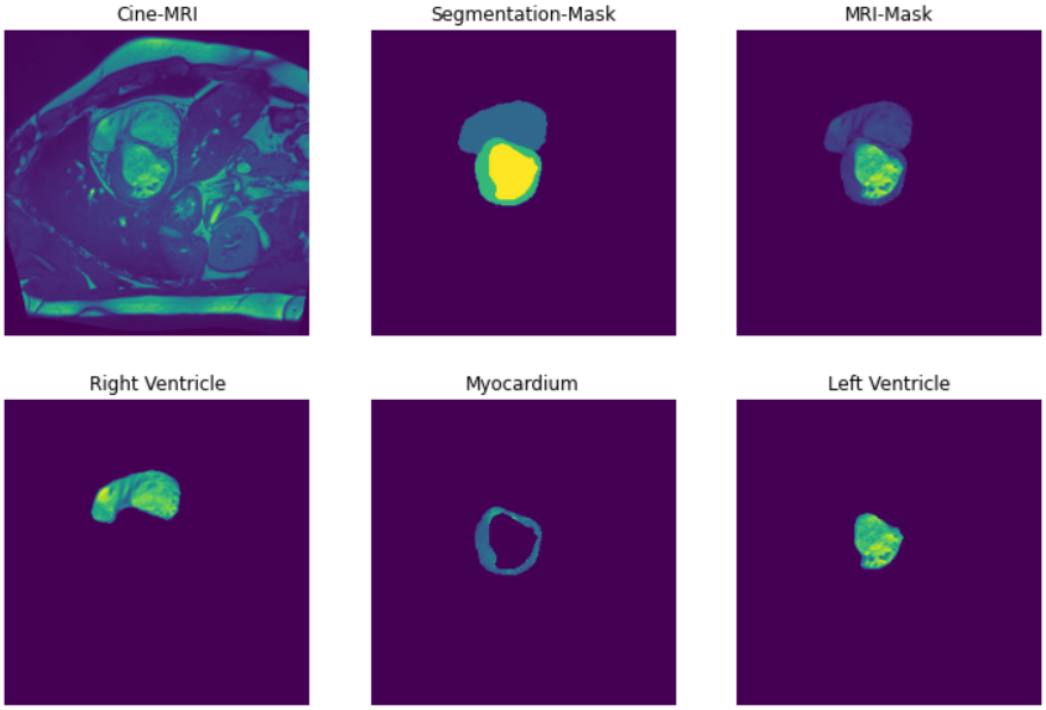

FIGURE 2.1: Images, Mask and heart structures

2.2 UK BioBank

UK Biobank (UKBB) [2] is a major international health resource with collections of data of 500,000 patients people among ages between 40-69 years in 2006-2010 who are periodically followed up for an expected period of 30 years from its start. These follow-ups collect different health-related variables of the volunteer; including genetic information, imaging of heart, brain, abdomen, bones and carotid artery, as well as biochemistry tests and questionnaires to characterize cognitive function and daily-life. The study has gathered complete information of patients with a wide range of diseases such as cancer, heart diseases, stroke, diabetes, arthritis, osteoporosis, eye disorders, depression and forms of dementia and has been used by many researches around the world with the aim of improving prevention, diagnosis and treatment (He and al, 2019; Alaa et al., 2019). In this study, we have selected patients with the following cardiovascular diseases: atherosclerotic heart disease, atrial fibrillation, angina, and hypertension. Next, we provide a short description [3] of the different groups used in the study:

[2] https://www.ukbiobank.ac.uk/
[3] https://medlineplus.gov/

- Atherosclerotic heart disease (**ATH**) : Atherosclerosis is a disease in which plaque builds up inside the arteries. Plaque is a sticky substance made up of fat, cholesterol, calcium, and other substances found in the blood. Over time, plaque hardens and narrows your arteries. That limits the flow of oxygen-rich blood to your body.
- Atrial fibrillation and flutter (**ATFB**): Atrial fibrillation or flutter is a common type of abnormal heartbeat. The heart rhythm is fast and most often irregular.
- Angina (**ANG**): Angina is a type of chest discomfort caused by poor blood flow through the blood vessels (coronary vessels) of the heart muscle (myocardium).We joined several patients with angina-related diseases as unstable angina and angina pectoris.
- Hypertension (**HYP**): High blood pressure is a common condition in which the long-term force of the blood against your artery walls is high enough that it may eventually cause health problems, such as heart disease.
- Healthy patients (**HEALTHY** Patients): Patients with no disease diagnosed.

There are two ways that diseases are assigned in the database: self-reported and diagnosed by a doctor. In table 2.2 is presented the number of patients in each disease subgroup and the origin of the disease. For the classification task, healthy patients were randomly selected to match the number of patients selected for each disease. In figure 2.2 we can observe an image of a patient for UKBB and the mask for ED and ES. As with the ACDC dataset, the manual segmentation of the region of interest (ROI) for images has been provided by the *Laboratory of Artificial Intelligence in Medicine* of the UB.

- Self Reported: Patients fill out a questionnaire pointing out the diseases they have been previously diagnosed with.
- Diagnosed by doctor: Diagnosis based on hospital inpatient admission and through linkages to a range of health-related records.

TABLE 2.2: Subset of Patients for each disease UKBB

Disease	Subset of Patients	Diagnose Modality
Angina	400	Subset combined
Atherosclerotic	500	Diagnosed by doctor
Atrial Fibrillation	220	Subset combined
Hypertension	780	Self Reported

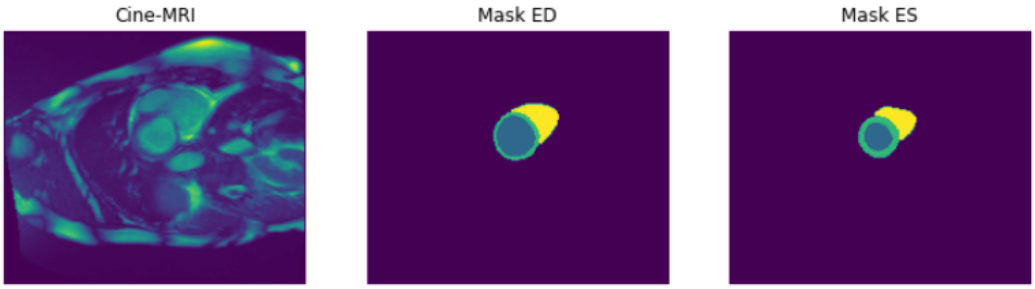

FIGURE 2.2: Cine-MRI and Masks for ED and ES

Chapter 3

Methods

This section details the methods used in the present study. A brief introduction to radiomics and their different types is provided. In addition, the adapted machine learning (ML) and deep Learning (DL) methods are presented. Finally, the innovation of this thesis, which consists of the fusion of the two types of radiomics generated is explained and an outline of the methodology adopted is shown.

3.1 Hand-Crafted Radiomics

The concept of radiomics is derived from the field of radiology inspired by the *-omics'* suffix, widely used in biology, describing the detailed characterization of biologic molecules including proteins, genes, among others (Cetin et al., 2017; Hosny, Aerts, and Mak, 2019). Following the same idea, radiomics is a method that consists in the extraction of a large amount of quantitative and qualitative information from medical images, with the aim of providing more information for the study of the physiopathology of tissue in the area of interest (Martin-Isla et al., 2020). Radiomics is used to obtain prognoses and predictive models. They have been used for studies of tumor regions, metastatic lesions, as well as in normal tissues. In this study, we will be using radiomics to obtain quantitative insights into heart structures to support predictive models of cardiovascular disease.

The principle underlying the use of radiomics and its increasing incursion into the field of research is that it is capable of providing sufficient information to generate support for the diagnosis and/or prognosis of the disease under study. Many times, the study of radiomics complemented with medical or biological information results in the generation of models with higher confidence and greater accuracy in prediction (Afshar et al., 2018). Radiomics can be divided into two main types: hand-crafted radiomic features (HCR) and deep learning-based radiomics (DLR). The latter will be presented in section 3.3.

3.1.1 Hand Crafted Radiomics Feature Extraction

The process for the extraction of HCRs consists in the following steps: (i) Image Acquisition, (ii) Identifying the volumes of interest , (iii) Manual or Semi-Automatic segmentation of the Volumes, (iv) Extracting and qualifying descriptive features of the volume and (v) Statistical analysis and model building. Figure 3.1 illustrates a representation of the steps aforementioned. For the extraction of these features we have used the open-source library *PyRadiomics* (Griethuysen et al., 2017)[1] using the

[1]Python Library PyRadiomics: https://pyradiomics.readthedocs.io/en/latest/

CMR and the respective masks of two cardiac phases, end-diastole (ED) and end-systole (ES), as shown in figure 2.2.

These large number of features are commonly categorized in three main groups: Shape, intensity, texture (Afshar et al., 2018). Next, we make a short description of the features extracted:

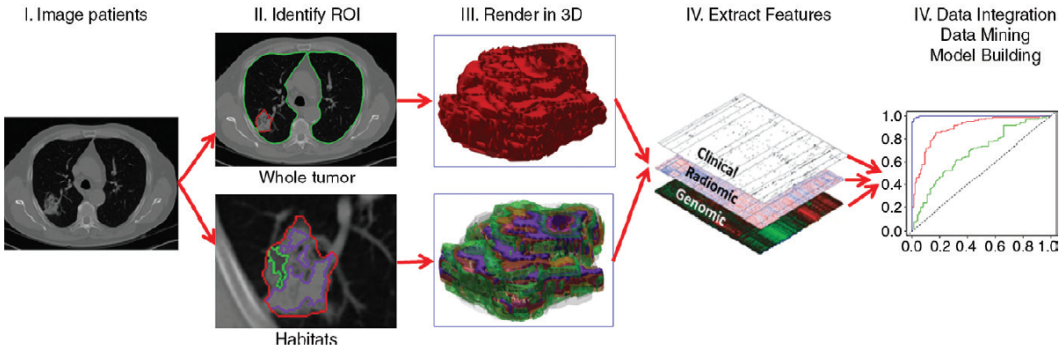

FIGURE 3.1: Radiomics Features Extraction Pipeline (Gillies, Kinahan, and Hricak, 2016)

- Shape Features: Quantification of geometric characteristics of the structures under study.
- Intensity Features (First Order): Statistics of the intensity distributions within the region of interest.
- Texture Features: Statistical texture refers to the stochastic or random properties of the spatial distribution of grey levels within an image using statistical measures, such as marginal probabilities.These second- or higher-order statistics are derived from grey-level intensity matrices, i.e. the CMR in this work.
  - Gray Level Co-occurrence (GLCM): Matrix that presents the number of times that two intensity levels have occurred in two pixels with specific distance.
  - Gray Level Run-Length (GLRLM): Matrix that presents the length of consecutive pixels having the same intensity.
  - Neighborhood Gray Tone Difference Matrix (NGTDM) : Quantifies the difference between a gray value and the average gray value of its neighbors within a predefined distance.
  - Gray-Level Zone Lenght Matrix (GLZLM): Considers the size of homogeneous zones in every dimension.
  - Gray-Level Dependence Matrix (GLDM): Quantifies the gray level dependencies in the ROI.

In the process of extracting HCR making use of the *PyRadiomics*, we have extracted 107 features, for each heart structure. Combining both phases, the total number of features entered to the ML pipeline is 642. In table 3.1 we can observe the number of feature of each category that were extracted.

TABLE 3.1: Number of features for each radiomics category

Category of Radiomics	Number of Features
First Order Statistics	18
Shape Based	14
Gray Level Coococurence Matrix	24
Gray Level Run Lenght Matrix	16
Gray Level Size Zone Matrix	16
Gray Level Dependence Matrix	14
Neighbouring Gray Tone Matrix Difference Matrix	5

3.2 Machine Learning

Machine Learning (ML) refers to statistical data-driven algorithms that are capable to learn the hidden patterns in a dataset and automatically generate an accurate prediction from that dataset. ML has been extensively applied in many industries, as in the biomedical field, in which since its beginning has helped researchers and clinicians with many important discoveries (Xiang et al., 2019; Woldaregay et al., 2019). ML does not provide the ground truth about a problem, but it has been adopted as a great tool in the decision-making for physicians for the quick and precise analysis it provides.

For this study, we have selected supervised learning methods to exploit the large number of features generated from the radiomics pipeline reftab: Radiomics Features and develop a classification model for the CVDs. Models such as support vector machines (SVM), random forests (RF) and logistic regression (LR) have been widely used in previous studies due to their well known performance (Nie et al., 2019; Bizzego et al., 2019).

Experiment Set Up

Towards the implementation of the ML pipeline and the classification of diseases, we have made use of *Scikit-Learn*[2] Python library, which is an open source library to develop ML at a high level of programming. In order for the models to perform better and faster, we have scaled our features to a range of 0 to 1 by making use of the *MinMaxScaler* from *Scikit-Learn*. Likewise, we have added to the ML pipeline a feature selection step based on *K-Best*. This filtering method works by scoring all the features based on univariate statistical tests and selecting the K best features. In this case, the tests selected have been both *Anova-F test* and *Chi*^{2} *test*. The training process has been set up to find the best combination of hyper-parameters of the models and the number of features among all the possible combinations. The decision of the best estimator is automatically selected by the best score among the average of accuracy scores of the cross validated groups. As for the classification algorithm, SVM has been chosen due to its flexibility and great presence in the different reference articles. Figure 3.3 presents an overview of the ML pipeline. For the optimization of the hyper-parameters, we have made use of the *Grid Search* too, also from the library. The hyper-pameter grid contents both univariate tests for the feature selection, a

[2]https://scikit-learn.org/stable/

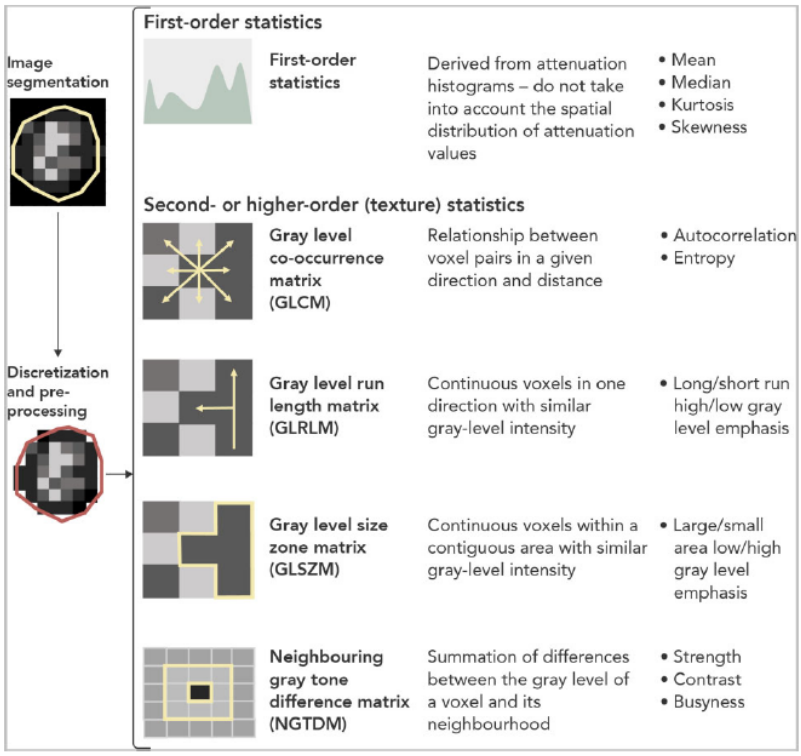

FIGURE 3.2: Radiomics Features (Oikonomou, Siddique, and Antoniades, 2020)

wide range of complexity numbers for the SVM and the kernel function to find the optimal hyperspace among linear function and radial-based function.

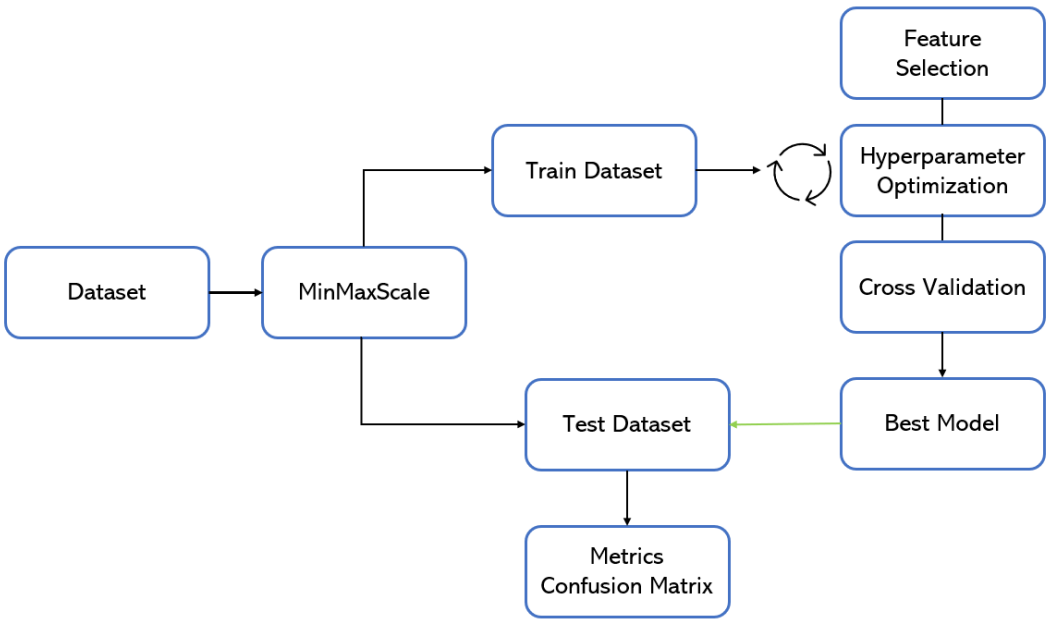

FIGURE 3.3: Machine Learning Pipeline Overview

3.3 Deep Learning-Based Radiomics

Medical image classification is an important area of research within the image recognition field. This task can be divided in two steps: (i) extracting the features from the raw image and (ii) using these features to create a relationship with the labels and make the classification (Lai and Deng, 2018). Deep learning-based radiomics (DLRs) are features extracted from the raw medical images via different architectures such as convolutions neural networks (CNN) and auto-encoders. In this study we focused on using CNN to extract high level features and we used different networks like Inception-V3, VGG-16, AlexNet Modules and Late Merging architectures (Nie et al., 2019). Some advantages of using deep learning to extract radiomics features include extraction features in an almost automatic fashion. The network can learn the features during the training process and then, extract learned features from unseen images. Another advantage is that almost the raw images can serve as input for the CNNs, since no segmentation of the area of interest (ROI) is needed.

We have used standard and self-designed CNN to extract the DLRs. It is highly important to decide which architecture to use to extract the radiomics and the quality of these is directly related to the type of architecture and its use. Below is a summary of the architectures used and information about them:

- InceptionV3: Widely known architecture for image recognition task, achieving a performance greater than 78.1% in the *Imagenet* dataset (Szegedy et al., 2015). The architecture is a combination of convolutional layers with batch normalization, average pooling, max pooling, concatenation of layers, dropouts and fully connected layers. This architecture holds 42 M parameters but is much more efficient than the VGG model. In figure 3.4 we show the base architecture of the InceptionV3.

- AlexNet Module: One of the first deep convolutional neural networks (DCNN) to participate in the *Imagenet* challenge showing a great performance (Krizhevsky, Sutskever, and Hinton, 2012). Figure 3.5 shows AlexNet based module. We performed two modifications separately to this architecture resulting in the networks AlexNet Module and AlexNet Dilated used in this work. For the first one, we reduced the number of neurons in the last fully connected layer, prior to the classification, to make it faster and reduce the number of parameters. For the second version, we maintained the original structure, but we added dilated component to the convolutions layers (Cui et al., 2019) . This dilation factor (DF) increases the space between original kernel elements resulting in an increased receptive field. We maintained the DF in 2. Figure 3.6 represents dilated convolutions with different DF.

- Late Merging: For this architecture, we have followed similar approaches as in (Nie et al., 2019; Bizzego et al., 2019). An overview of the late merging architecture used is provided in figure 3.7. This architecture consists in creating two identical convolutional branches merged in a fully connected layer. An important difference with respect to the reference architectures is that they have used 3D convolutions, which generates more information for the model and presents better results. Cine-MRI images are anisotropic volumes and, thus, the adaptation of a 3D model is requires different approaches out of the scope of this study (Liu et al., 2018). Therefore, this approach was dismissed and a

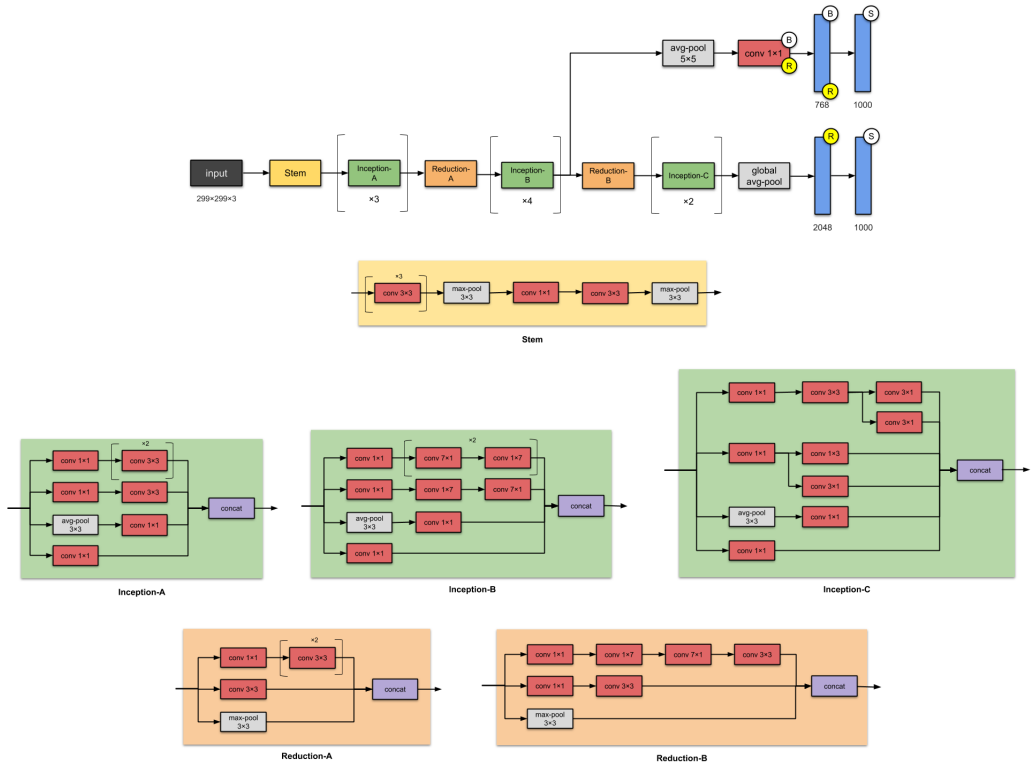

FIGURE 3.4: InceptionV3 Architecture (Karim, 2019)

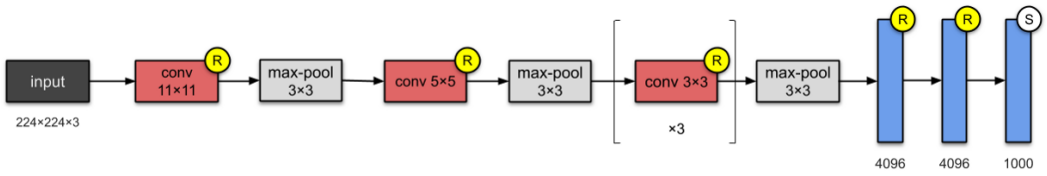

FIGURE 3.5: AlexNet Base Module (Karim, 2019)

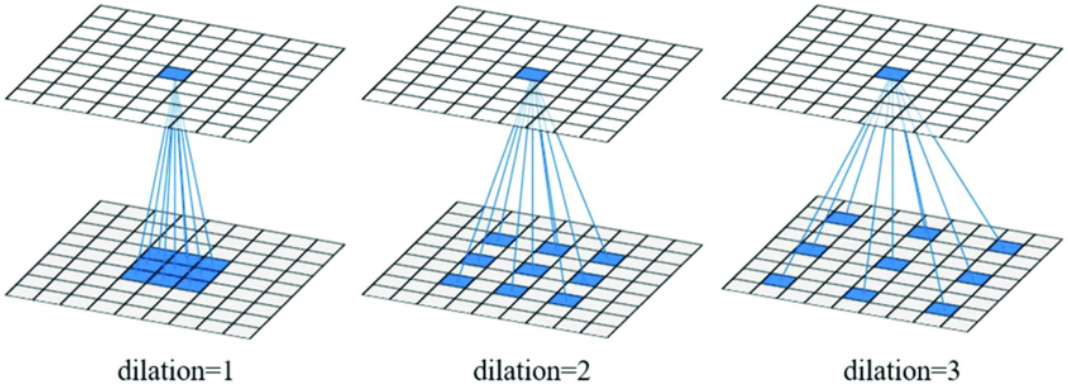

FIGURE 3.6: Dilated Convolutions (Cui et al., 2019)

multi-slice model was considered instead. The way we have used this architecture is to introduce each heart structure into separate convolutional branches and then combine all the information in a fully connected layer. In figure 3.8 we show the late merging model.

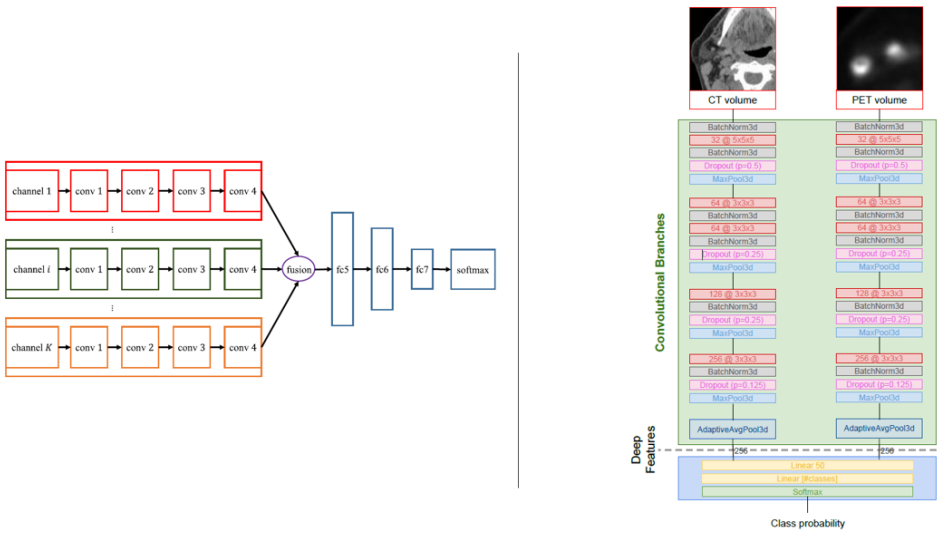

FIGURE 3.7: 3D Late Merging Architecture. Left (Nie et al., 2019). Right (Bizzego et al., 2019)

Experiment Set Up

To implement, load and train the CNNs and the trained models, we have made use the Python's library *TensorFlow* [3] and its API *Keras* [4]. For the training process of these architectures, data augmentation techinques such as flipping, rotation and zoom-in were used on the images to feed the network with a greater amount of images which helps to improve the performance and reduce over-fitting. *Categorical Cross-Entropy* was used as a loss function and the algorithm to update the weights in the training was *Stochastic Gradient Descent* (SGD).

3.3.1 Image Preprocessing

Medical images need to be pre-processed before they can be provided as input for CNN. The CMR slices had different sizes along x-y. Thus, we resized them to a common size: 150x150 voxels. The images also came with different number of slices for each patient, from 6 to 11 slices. Regarding this information, we decided to feed the networks with two different images prepossessed formats: with only middle-slice format and mult-slice format and evaluate each performance and complexity. For the Middle-slice Format, we used only the middle slice of each phases (End Diastole and End Systole). To build the input images, we used the multiplication of the CMR image and corresponding mask for each cardiac structure.The heart structures were stacked in the third dimension, having in total 3 channels. For the Multi-slice format, the middle and the neighbor slices were selected. The final arrays for each patient

[3] https://www.tensorflow.org/
[4] https://keras.io/

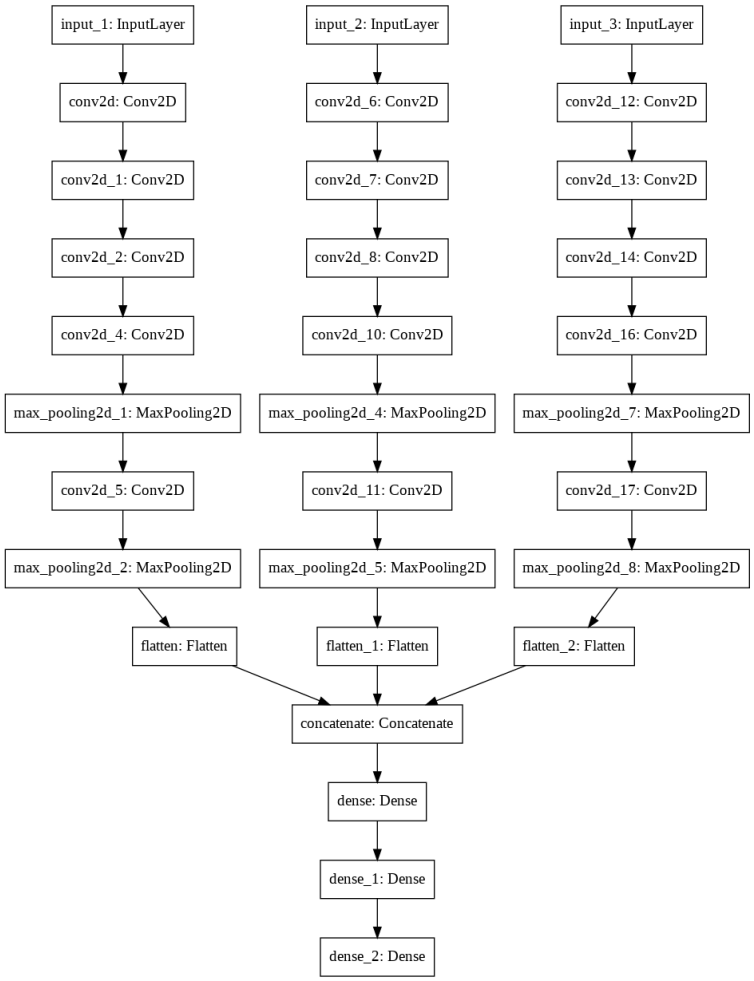

FIGURE 3.8: Late Merging Architecture

were stacked vertically. It was assumed that this type of data augmentation might provide more information to CNN and improve the results, aware of the increased complexity of the model that this entails. Figure 3.9 illustrates a representation of how the images were formatted to be fed to the CNN.

For the UKBB, the images received an enhancement by the *Histogram Matching* technique. The process consists of matching the intensity histogram of an image set as template and then transforming the following images' histogram matching the template's histogram. Figure 3.10 shows an example of this technique.

3.3.2 DLR Extraction

From the training, CNNs are capable to learn underlying features from the image dataset through its filters. These features create the mapping of the input sample with the corresponding label. When s CNNs are trained, the new inputs activate certain neurons that ultimately correspond to the correct label of the input. Those features are concentrated in the last layer before the classification layer. Finally, this last layer is where classification takes place, usually activated with the *softmax* function. Those underlying features are the DLR used in this study. For the extraction, after the model is trained, we extract the features for each patient by removing the

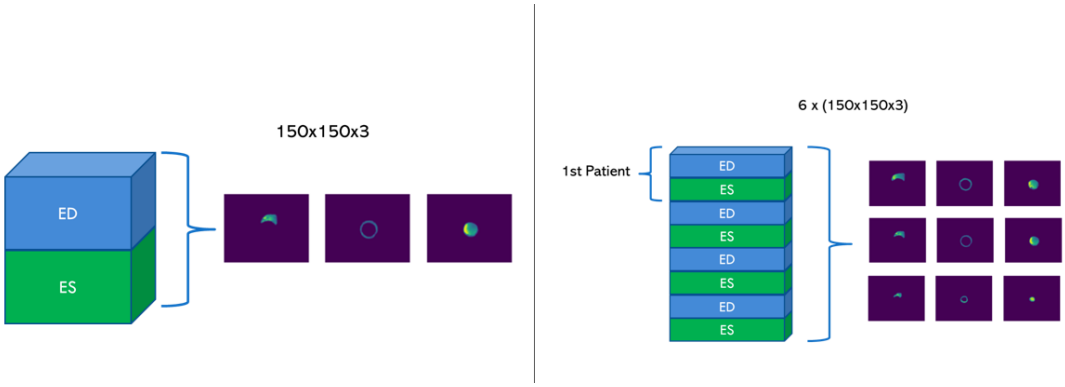

FIGURE 3.9: Images Format for input to the CNNs

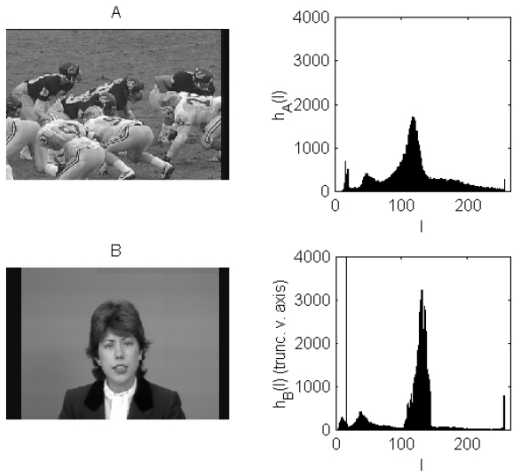

FIGURE 3.10: Histogram Matching Example (Maini and Aggarwal, 2010)

classification layer. The output vector will contain the learned features for each patient. In table 3.2, we provide the number of features contained in the output vector of the different CNNs used.

3.4 Integration of Features

In the previous section we have provided a brief introduction to radiomics and how they can be generated from raw images. Both methods have advantages and disadvantages (Afshar et al., 2018). These are summarized in the table 3.3. The innovation of this thesis is the fusion of the two types of radiomics. This fusion, as seen in 1 section 1.1, has been implemented for tumor detection, cancer diagnosis, survival time. However, to the best of our knowledge, it has not been performed in the analysis of CVDs. Using this fusion scheme, we aim to is to capture different and complementary information from the input images, providing insightful information for the predictive model to perform better.

There are two common approaches for fusion of these radiomics (Afshar et al., 2018): *Decision Level Fusion* and *Feature-Level Fusion*. The second is the method selected in

TABLE 3.2: Number of output vector features of CNN

Model	Number of Features
Inception	2068
AlexNet Module	5
AlexNet dilated	4096
Late Merging	256

this thesis, which consists on the concatenation of the features to then feed them to the classifier. For this method, we combine the radiomics pipeline and deep learning feature extraction pipeline to obtain our final results. Figure 3.11 illustrates the fusion of features pipeline developed.

TABLE 3.3: HCR vs DLR: Advantages and disadvantages.

HCR	DLR
Needs a prior knowledge on types of features to extract	Can learn features on its own and without human intervention
Features are typically extracted from the segmented ROI	Does not necessarily require segmented input
It is generally followed by a feature selection algorithm	Feature selection is rarely performed
As features are defined independent from the data, do not require big datasets	Requiere huge datasets, since it has to learn features from the data
Processing time is not normally significant	High computational cost depending on the architecture and size of the dataset
Since features are pre-designed, they are tangible	The logic behind the features and decisions is still a black box

3.5 Metrics

Throughout this study, we used different metrics to evaluate the model performance. For ACDC and multi-class problem, we focus only on accuracy (ACC) and on the confusion Matrix (CM). Regarding the experiments using the UKBB dataset and the binary classification problem, we add area under the receiver operating characteristic curve (ROC AUC), precision and recall. In table 3.4, we provide a summary of the metrics used in the study. To calculate these metrics, we used the *Metrics* module from *Scikit-Learn*.

TP: True Positives; TN: True Negatives; FP: False Positives; FN: False Negatives.

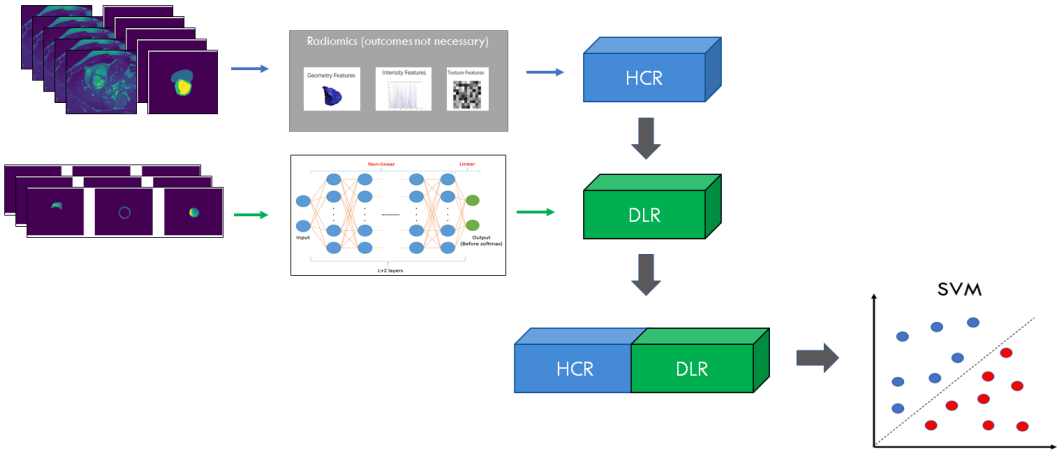

FIGURE 3.11: Fusion of Features Pipeline

TABLE 3.4: Description of metrics used in the study to evaluate model performance in classification.

Metric	Description
Accuracy	Measures the percentage of the algorithm classifying the input data correctly $\frac{TP+TN}{TP+TN+FP+FN}$
Confusion Matrix	Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class
AUC ROC Curve	Measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.
Precision	The ability of the classifier not to label as positive a sample that is negative $\frac{TP}{TP+FP}$
Recall	The ability of the classifier to find all the positive samples $\frac{TP}{TP+FN}$

Chapter 4

Results

4.1 ACDC Dataset

In this section, we present the results obtained in the ACDC Dataset. Section 4.1.1 corresponds to the results obtained by the ML Pipeline with HCR. Section 4.1.2 includes the results of the classification in the test dataset by means of the CNN architectures presented in Chapter 3. Lastly, we provide the results obtained using the methodology for fusion of features detailed in section 4.1.3. Results after the integration of medical information are not presented because in any of the experiments where they were considered, either height or weight were selected by the algorithm.

4.1.1 Machine Learning

During the extraction of the HCR with *PyRadiomics*, we end up with 387 features for each phase, having a total of 774 for both phases. Some of the features are pure informative from the library or some are extracted by already deprecated. The final number of features introduced to the pipeline are shown in table 3.1 in section 3.1.

Following the pipeline presented in Chapter 3, we use SVM as the ML learning model for the classification of the diseases. The Grid Search parameters comprehended many values for the complexity parameter and also the kernel type for the separability. The training set was divided in four stratified sets for cross validation and the best estimator was chosen as the one with the best average accuracy among the cross validation.

In figure 4.3, we present the confusion matrix of the results in the test dataset with the best estimator obtained from the pipeline. In figure 4.2 are provided the top 15 feature selected by the algorithm in order of importance for the final classification. We can observe there is a wide variety of categories of radiomics being selected by the algorithm, being the surface volume ratio, kurtosis and gray level top 3 in importance. Also, it is interesting to notice that for the top features presented, there is an important presence of features corresponding to the ES phase. From the total of 88 features selected by the algorithm, 55 correspond to the ES phase and 33 to the ED phase. This can lead to the intuition that the ES phase can provide more predictive power to the algorithms.

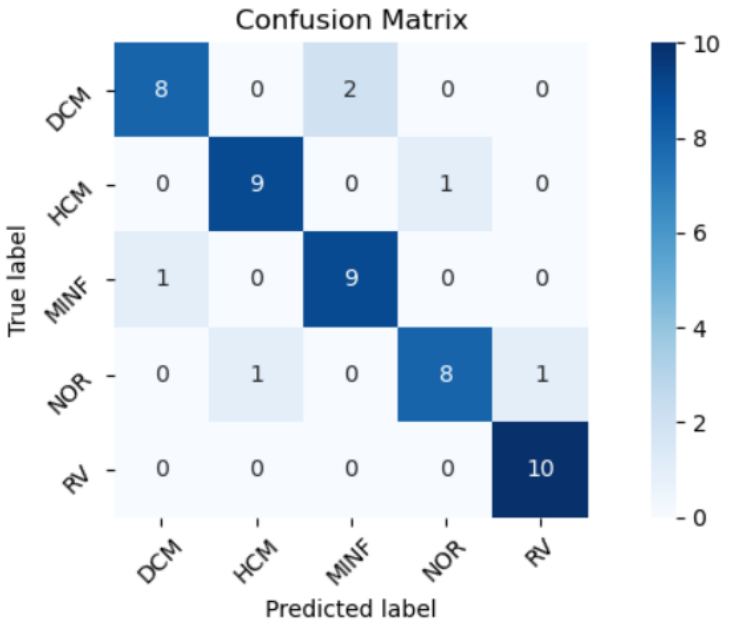

FIGURE 4.1: Confusion Matrix of the Test Dataset

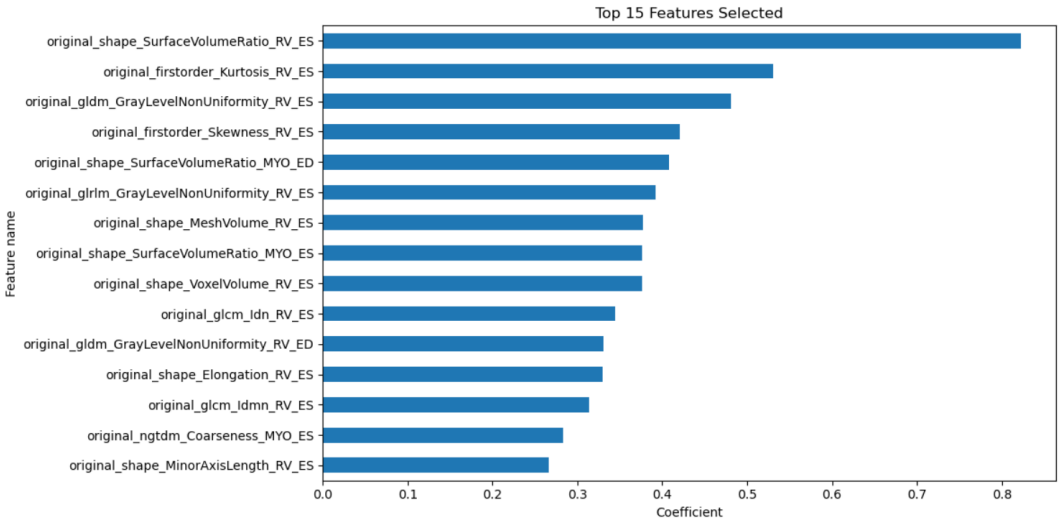

FIGURE 4.2: Top 15 Features

4.1.2 Deep Learning

For this section, we used the mentioned models AlexNet modules, Inception V3 and Late Merging Model. In tables 4.1, 4.2, 4.3 we present the results of the CNN performances for Middle Slice, Multi-slice performance in ES dataset and ED dataset, respectively. For the Multi-Slice format, The accuracy was calculated by a majority vote among the predictions of the 6 samples (slices) of each patient in the test. For the Middle-Slice format, the predictions were calculated for the test sets separately for both ES and ED phases.

From the tables, we can observe that we obtain better results for the models trained

using only the middle slice. The model that presented the best results correspond to the AlexNet Module with a feature vector of 5 features, with an accuracy of 0.86. Also, we can observe that ES phase set, present better results compared to the ED set.

TABLE 4.1: CNN Performance Multi-Slice

Model	Metric	DCM	HCM	MINF	NOR	RV	Accuracy
Inception	Precision	1	0.86	0.47	0.80	0.75	0.70
	Recall	0.70	0.0.60	0.0.90	0.40	0.90	
	f1-score	0.82	0.71	0.62	0.53	0.82	
AlexNet Module	Precision	0.89	0.89	0.6	0.4	0.86	0.70
	Recall	0.8	0.8	0.9	0.4	0.6	
	f1-score	0.84	0.84	0.72	0.4	0.71	
AlexNet Dilated	Precision	0.73	0.90	0.58	0.75	0.89	0.76
	Recall	0.80	0.90	0.70	0.60	0.80	
	f1-score	0.76	0.90	0.64	0.67	0.84	
Late Merging	Precision	0.64	1.00	0.64	0.75	0.91	**0.78**
	Recall	0.70	0.90	0.70	0.60	1.00	
	f1-score	0.67	0.95	0.67	0.67	0.95	

TABLE 4.2: CNN Performance Middle Slice Test Set End Systole

Model	Metric	DCM	HCM	MINF	NOR	RV	Accuracy
Inception	Precision	0.53	0.77	1.00	0.71	0.90	0.70
	Recall	1.00	1.00	0.1	0.50	0.0.90	
	f1-score	0.69	0.87	0.18	0.59	0.90	
AlexNet Module	Precision	0.83	0.77	0.88	1.00	0.91	**0.86**
	Recall	1	1	0.7	0.6	1	
	f1-score	0.91	0.87	0.78	0.75	0.95	
AlexNet Dilated	Precision	0.69	0.71	0.5	1	0.91	0.72
	Recall	0.9	1	0.5	0.2	1	
	f1-score	0.78	0.83	0.5	0.33	0.95	
Late Merging	Precision	0.75	0.77	0.71	0.8	0.77	0.76
	Recall	0.9	1	0.5	0.4	1	
	f1-score	0.82	0.87	0.59	0.53	0.87	

4.1.3 Fusion of Features

In tables 4.4 and 4.5, we provide comparative of results of the combinations with the DLR extracted by means of the different architectures. As can be appreciated, with the combination of radiomics, the results have not improved the benchmark obtained by HCR only. The best combination in our set of experiments is with the AlexNet Module DL model. In figure 4.4, we provide the confusion matrix of the best result. From the CM, we can observe that the diseases are all correctly predicted

TABLE 4.3: CNN Performance Middle Slice Test Set End Diastole

Model	Metric	DCM	HCM	MINF	NOR	RV	Accuracy
Inception	Precision	0.53	0.91	0.33	1	0.91	0.70
	Recall	1	1	0.2	0.3	1	
	f1-score	0.69	0.95	0.25	0.46	0.95	
AlexNet Module	Precision	0.71	0.9	0.6	0.83	0.9	**0.78**
	Recall	1	0.9	0.6	0.5	0.9	
	f1-score	0.83	0.9	0.6	0.62	0.9	
AlexNet Dilated	Precision	0.64	0.82	0.45	1	0.91	0.72
	Recall	0.9	0.9	0.5	0.3	1	
	f1-score	0.75	0.86	0.48	0.46	0.95	
Late Merging	Precision	0.64	0.83	0.5	0.75	0.83	0.72
	Recall	0.9	1	0.4	0.3	1	
	f1-score	0.75	0.91	0.44	0.43	0.91	

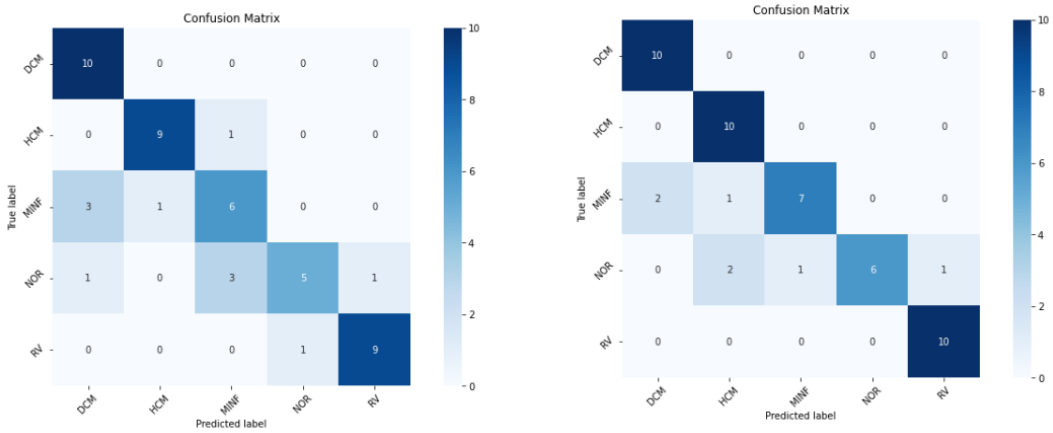

FIGURE 4.3: Confusion Matrix of the Test Dataset with AlexNet Module (ED on the right, ES on the left)

TABLE 4.4: Fusion of Features Result Multi-Slice Format

ML Model	DLR Model	Modality	Accuracy
SVC	Inception	HCR+DLR	0.72
SVC	AlexNet Module	HCR+DLR	0.88
SVC	AlexNet Dilated	HCR+DLR	0.78
SVC	Late Merging	HCR+DLR	0.58

and only the Normal patients label present false negatives. The feature selection algorithm used a combination of 5 HCR and selected the 10 DLR features. In Figure 4.5 is shown the importance of these features for the classifier and we can notice that DLR features corresponding to the activation of the ES and ED dataset are in the highest position in terms of importance to make the classification. Followed by a combination of HCR and DLR. Also, from comparing these features with the top 15

TABLE 4.5: Fusion of Features Result Middle Slice Format

ML Model	DLR Model	Modality	Accuracy
SVC	Inception	HCR+DLR	0.78
SVC	AlexNet Module	HCR+DLR	**0.88**
SVC	AlexNet Dilated	HCR+DLR	0.84
SVC	Late Merging	HCR+DLR	0.84

features from figure 4.2, we conclude that mesh volume, voxel volume and minor axis length for the RV and ES repeat.

We observe a reduction in the false negatives and all the diseases are predicted correctly. This result, despite not improving the previous accuracy, it is important in the field of medicine since by classifying diseases correctly and presenting false negatives when the disease does not exist generates more attention in the case of a hypothetical positive diagnosis (no disease), on the contrary, it would generate little attention in cases where the diagnosis is relevant to treat the disease correctly and in time. This specific pattern of model performance could be specific to the random seed that generated the data split and model parameters. In Appendix A we repeated this experiment in with several seeds and found that the pattern repeated across different seeds. This result suggests that HCR+DLR robustly reduces false negatives.

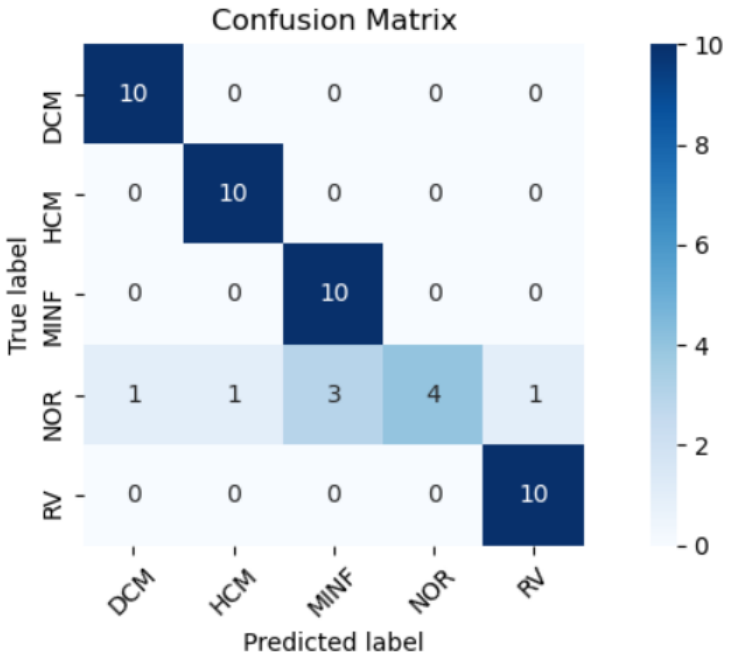

FIGURE 4.4: Confusion Matrix of the Test Dataset Fusion of Features

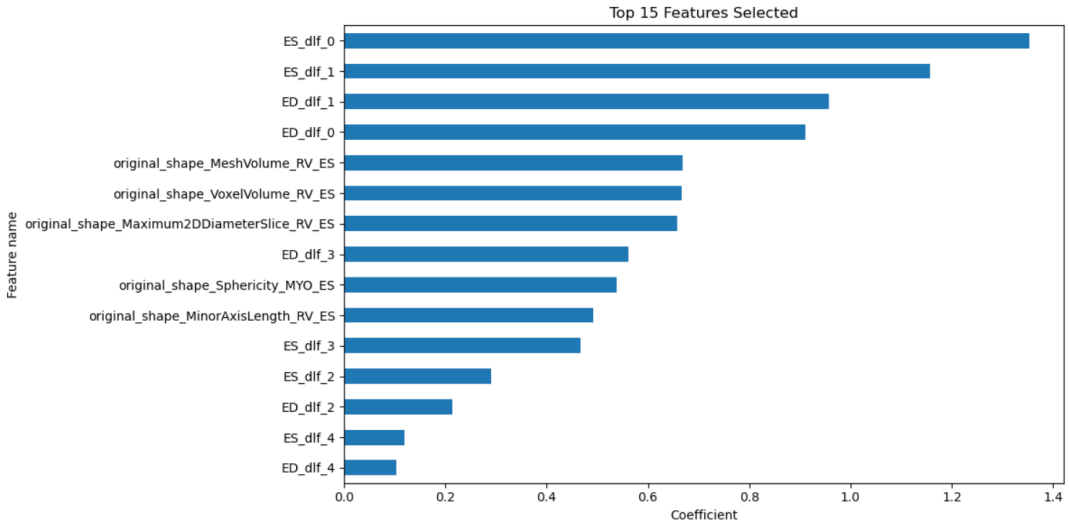

FIGURE 4.5: Chosen Features

Comparison with Benchmark

Reference (Bernard et al., 2018) shows the Top5 results for ACDC classification challenge. Figure 4.6 shows the table of results. In comparison to the results obtained in this study, with an accuracy of 0.88, HCR alone have performed within the range of these top five results. This shows HCR are a powerful tool to make predictions with medical images. Nonetheless, this methodology was not able to outperform the best result obtained for this dataset, which was 0.96 (Khened, Varghese, and Krishnamurthi, 2019) by learned features from segmentation model and using ensemble classifiers. We believe more powerful techniques for feature selection and a larger exploration of CNNs can lead to higher accuracies.

TABLE VI
RESULTS ON THE CLASSIFICATION CHALLENGE.

Methods		***Accuracy***
Authors	Architectures	
Khened *et al.* [46]	Random Forest	**0.96**
Cetin *et al.* [53]	SVM	0.92
Isensee *et al.* [44]	Random Forest	0.92
Wolterink *et al.* [50]	Random Forest	0.86

FIGURE 4.6: ACDC Challenge Top5 Classification

4.2 UK BioBank

We conducted a similar analysis using UKBB as with the previous dataset. However, for UKBB, we solved a binary classification problem, instead of a multi-class. Multi-class approach was discarded because, unlike the ACDC dataset, the groups are not as well defined, with respect to diseases. Nonetheless, we still made a first approach to the multi-class problem. However, as expected, the results were low

and inconclusive, and further work will be needed. Multi-class results are added in the appendix B.

4.2.1 Machine Learning

As explained before, in this section we show the results of the performance of the ML algorithm with only HCR. In table 4.6 we provide a summary table and in the following figures the CM and AUC of each disease.

In the table we find an acceptable class separation for angina, atherosclerotic and hypertension with 0.77, 0.83 and 0.73 accuracy, respectively. For atrial fibrillation we found a 0.6 which shows more difficulty for the model to find a greater separability in the data. One explanation for this may be that at the image level, patients with this disease do not present a great difference with respect to patients and their diagnosis can be found with the help of other tests such as electrocardiograms [1], among others.

TABLE 4.6: HCR Results UKBB

Disease	Accuracy	AUC ROC	Precision	Recall
ANG	0.77	0.8372	0.728	0.86
ATH	0.82	0.8996	0.7962	0.86
ATFB	0.6	0.63812	0.5833	0.77
HYP	0.73	0.836	0.7513	0.697

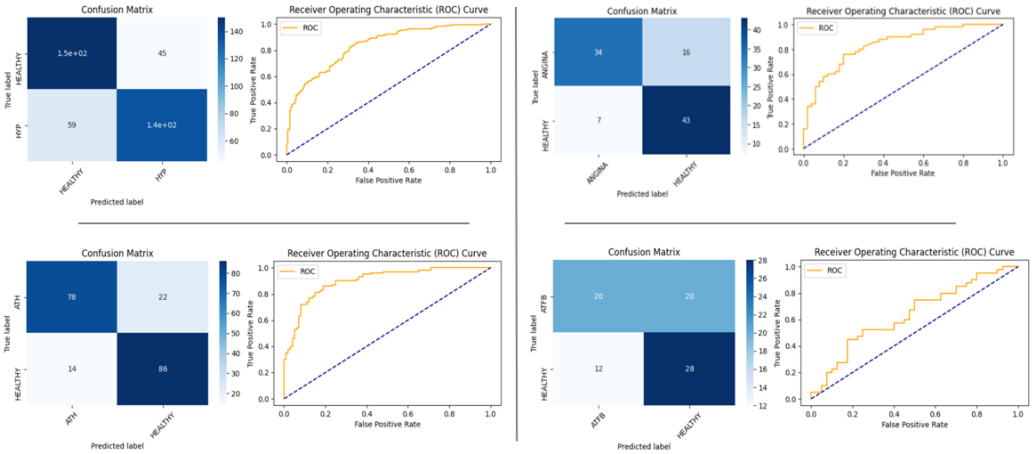

FIGURE 4.7: Confussion Matrix and ROC AUC showing the performance of HCR for each disease

4.2.2 Deep Learning

For Deep Learning classification, given that for UKBB the datasets were larger in size, we decided to train the models only with the middle slice format 3.9, since for

[1] https://www.mayoclinic.org/diseases-conditions/atrial-fibrillation/diagnosis-treatment/

ACDC we did not perceive many advantages in performance with respect of training using the middle slice only. In table 4.7, we show the results for the DL testing in the portion of the test set and the accuracy for the complete dataset for both ED and ES phases. We can find the best performers for each disease in bold. As for the previous dataset, the DL performance has not reach the HCR results.

TABLE 4.7: Deep Learning Results UKBB

Disease	Model	ACC	AUC ROC	Precision	Recall	ACC ED	ACC ES
ANG	AlexNet	0.58	0.657	0.545	0.96	0.544	0.625
	InceptionV3	**0.692**	0.754	0.697	0.68	0.725	0.675
	Late Merging	0.5725	0.52	0.547	0.547	0.576	0.566
ATH	AlexNet	0.716	0.784	0.690	0.79	0.710	0.739
	InceptionV3	**0.720**	0.801	0.713	0.737	0.716	0.739
	Late Merging	0.560	0.565	0.539	0.813	0.51	0.6
ATFB	AlexNet	0.55	0.523	0.61	0.6	0.597	0.602
	InceptionV3	0.535	0.554	0.523	0.81	0.541	0.55
	Late Merging	**0.602**	0.592	0.66	0.432	0.60	0.54
HYP	**AlexNet**	**0.687**	0.751	0.724	0.605	0.69	0.71
	InceptionV3	0.653	0.701	0.70	0.543	0.628	0.706
	Late Merging	0.570	0.627	0.563	0.626	0.592	0.542

4.2.3 Fusion of Features

For this section, it was decided to take the best DL model to perform the fusion with HCR. Since UKBB are larger, training times would take longer. Also, we have limited the number of features in the grid search to only 1000 features, because in some cases, we can have a feature vector of up to 8000 features, which would generate a high level of overfiting and would not generate generalizable predictions. In the table 4.8, we can observe that the merge with DLR does not present important improvements with respect HCR. This may have to do with the low accuracy found in the DL performance. We also added a bar plot with the performances for each modality in figure 4.8.

TABLE 4.8: HCR+DLR Results UKBB

Disease	Accuracy	AUC ROC	Precision	Recall
ANGINA	0.73	0.8048	0.717	0.76
ATHEROSCLEROTIC	0.79	0.822	0.8085	0.76
ATRIAL FIBRILIATION	0.6	0.619375	0.5869	0.675
HYPERTENSION	0.736	0.788	0.75	0.708

Summary of Results for UKBB

TABLE 4.9: Summary of Accuracy for each method UKBB

Disease	HCR	DLR	DLR+HCR
ANGINA	0.77	0.692	0.73
ATHEROSCLEROTIC	0.82	0.72	0.79
ATRIAL FIBRILIATION	0.6	0.602	0.6
HYPERTENSION	0.73	0.682	0.73

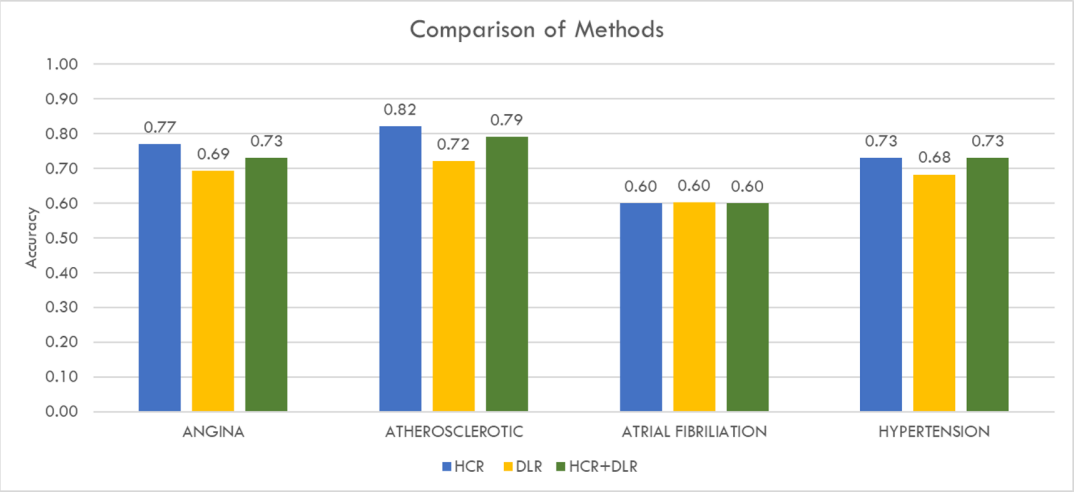

FIGURE 4.8: ML methods comparison UKBB

Chapter 5

Discussion

5.1 Conclusions

In this work, we have performed an extensive study of data mining from medical images. We have developed a pipeline for the extraction two types of: Hand Crafted Radiomics (HCR) versus Deep Learning-based Radiomics (DLR) and compare the performance of classification for each of the methods, demonstrating the ability of both to accurately diagnose between patients with cardiovascular diseases and healthy patients from medical images. Likewise, we performed a fusion for both types of features, with the aim that both features set would provide complementary information to the classifiers and, hence, obtain better results. Following we show a summary of the best findings in this study:

- HCR perform equally well or outperform DLF
- In almost all cases, the fusion of HCR+DLR resulted in selecting a combination of the two types of features as the most informative.This fact demonstrates that both types of radiomics share importance in the final decision of the classifier
- We demonstrated that CMR contains valuable information that can be extracted using either human-engineered, deep learning or both approaches to allow accurate CVD diagnosis.

In this thesis, several Machine Learning (ML) and Deep Learning (DL) concepts were applied to generate an innovative method for the fusion of radiomics for classification of cardiovascular diseases (CVDs). To the best of our knowledge, this is the first study to explore both hand-crafted and deep learning-based radiomics for the classification of CVDs.

5.2 Challenges

Many challenges were found during the study and the development of the pipeline for the extraction of the models. Specifically, CNNs were trained with only one or three slices. Adding more slices can entail more information about the volume in study. Nevertheless, more slices require more computation and the CNNs need to be carefully designed to manage such information. Also, there are other powerful options for feature selection techniques available that can return a higher accuracy. The following section we list applications to the mentioned limitations which must be addressed in future work to obtain stronger conclusions.

5.3 Future Work

In this section we elaborate on different approaches that were considered during the realization of the project, but for many different reasons, could not be part of the thesis.

5.3.1 Different Type of Fusion

In 3 we explained that the fusion was performed in a *Feature Level* fashion, where the features were concatenated and then classified with a SVM machine, similar procedure taken by (Nie et al., 2019). Another way to perform the classification of the features, consists in creating a multi-input Neural Network model, combining all the information for each type of data in a fully connected layer. In figure 5.1 is shown an example of this approach. This architecture, despite presenting a great challenge since the set up can be difficult when working with two types of data, as could be numeric/continuous data and images, can return a great performance for the complexity of the model itself.

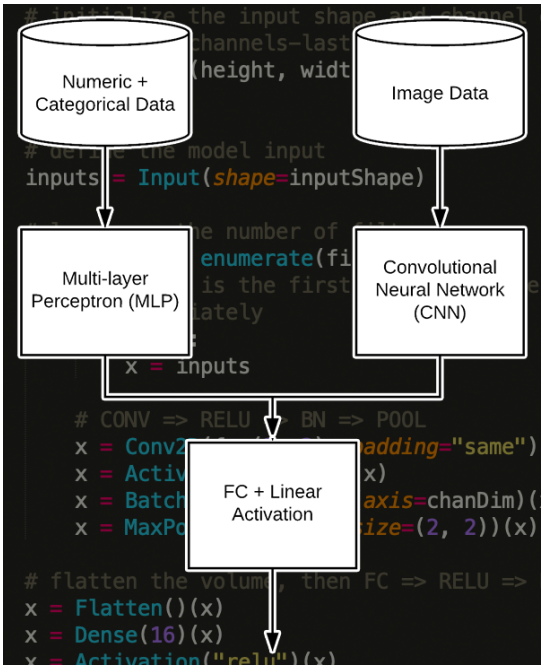

FIGURE 5.1: Multi-Input NN

5.3.2 Computer Vision Extensive Study

As we mentioned through this report, Computer Vision is a large area of study with many strategies for prepossessing the images and also, many available architectures. For this point, we considered training the models with pre-load weights, as for the *Imagenet*. This approach could have saved training time and it could have returned better results. Also, we considered many image recognition architectures, as for example: *VGG-16*, *Dense-Net*, *ResNet-50* and many others. We performed some initial experiments with these mentioned models, but we segmented our research to the models explained in 3 section 3 following the best results obtained. Also, we observed having large quantity of features in the output vector does not provide good

results. One solution could have also modify these networks and add a smaller output vector before the classification layer.

5.3.3 Correlation Study Feature

It has been explained that the aim of the fusion of both types of radiomics, HCR and DLR, is to provide complementary information one to another. Nevertheless, of the large number of features that can be extracted from both categories, many of them can be present be correlated. In (Bizzego et al., 2019), they performed a correlation analysis and features with high correlation were removed. Likewise, it was considered to perform the research of correlation of the features selected by the *Feature Selection* algorithm, obtaining an in-depth knowledge of the model selected.

5.3.4 Histogram Equalization

Histogram Equalization is a technique for image enhancement, widely used in the prepossessing stage for images prior to image recognition tasks. This image enhancement technique improves the intensity of the images, therefore, improving the algorithm's accuracy for the radiomics extraction. Figure 5.2 shows an example of histogram equalization for medical images prior feature extraction for cancer classification (Patel, BharathK., and Muthu, 2020).

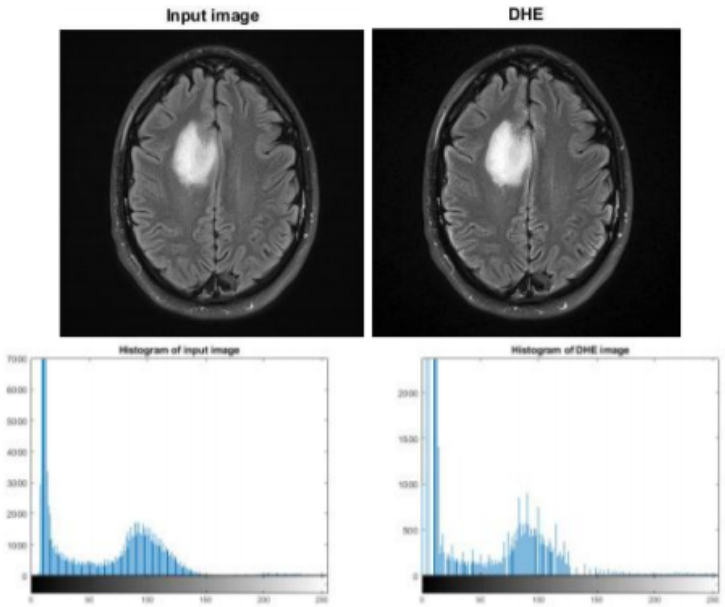

FIGURE 5.2: Histogram Equalization Medical Images (Patel, BharathK., and Muthu, 2020)

5.3.5 Interpretability

Deep Learning models present an interpretability challenge because the models are well known to be 'black-boxes'. This area of DL and Computer Vision is still in large development and research. As we developed the pipeline of the fusion of features as figure 3.11 shows, we found interest in understanding the DLR features selected by

the algorithm. There are several approaches to access to these features and develop an intuition of these features as presented (Moreira et al., 2020)

5.4 Code and Support Files

Jupyter notebooks, python scripts and support files can be found in the following Github page : Github Repository

Appendix A

ACDC Fusion of Features

A.1 Fusion of Features pattern

In 4 section 4.1.3 we showed how the CM for the fusion of features 4.4. This result, as we mentioned, this result is favorable in the field of medicine because in case of incorrect predictions it is preferred for favorable results, as it is in this case, that only false negatives occur for normal patients. In this line, we wanted to check that this pattern is maintained, changing the random factor in the model. Therefore, we repeated the experiment by changing the random seed for the selection of samples in the cross validation groups. In the figures A.1 and A.2, two confusion matrices are shown with the results varying that variable. We can see that in both cases, the pattern is maintained by qualifying, almost in its entirety, the diseases correctly. For this case, it was considered to make an experiment only with the diseases, however, in order to maintain the nature of the challenge, this experiment was rejected.

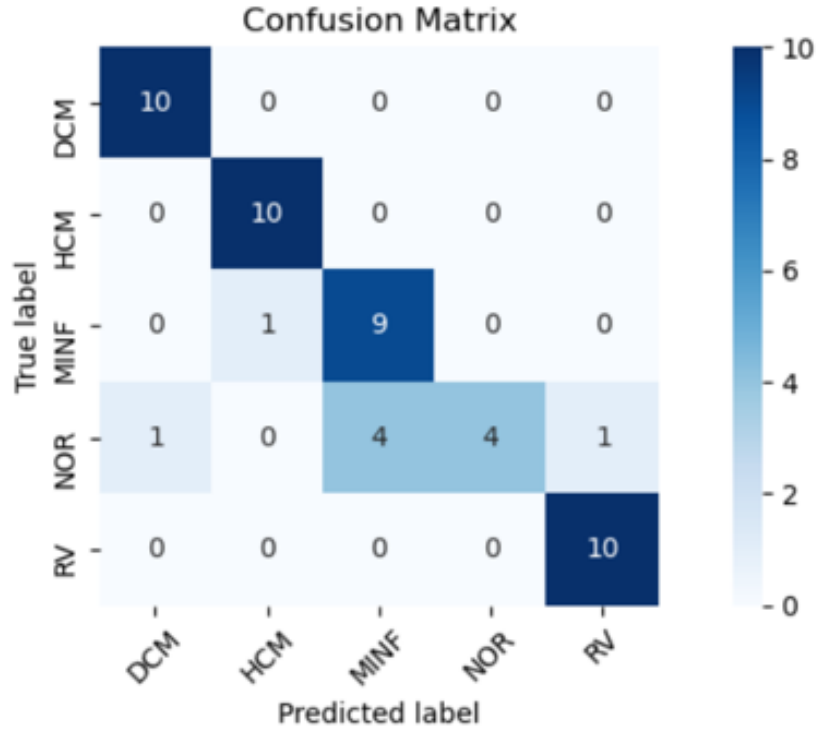

FIGURE A.1: Random Seed 1

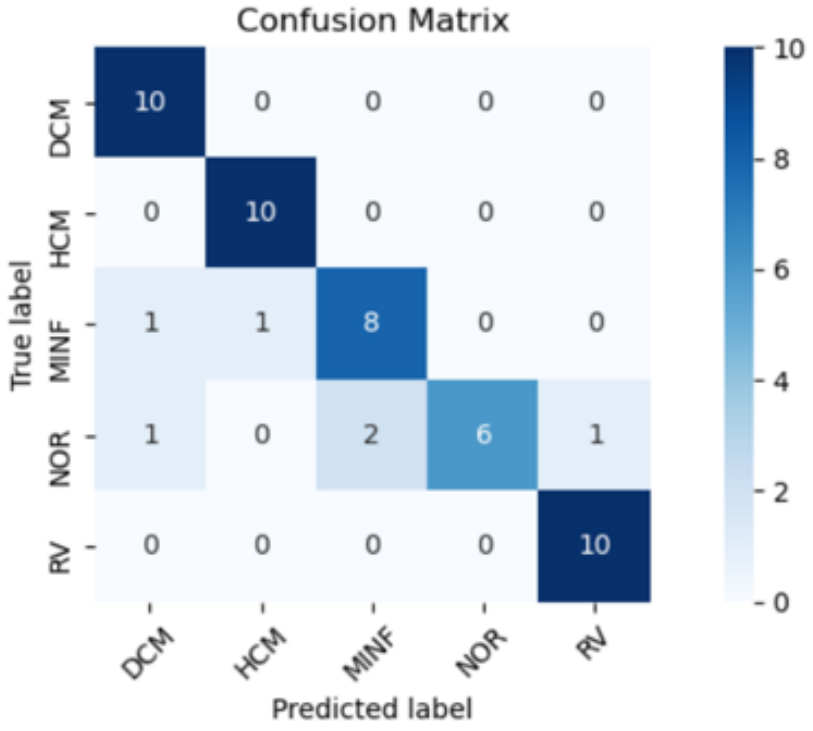

FIGURE A.2: Random Seed 2

Appendix B

UK Bio Bank Multi-class

B.1 UKBB Multi-Class Approach

It was mentioned in 4 section 3.2.1 that a first approach was tried to classify several diseases into a multi-class problem. However, it has been mentioned that, unlike the ACDC, the groups are not so well defined. There are even patients suffering from several of the diseases that were tried to be classified at the same time. Our first approach was not successful and although some work was done on the data, e.g. the patients with the same diseases were removed, the maximum precision obtained was 0.45. This result does not allow conclusions to be drawn and shows that for the model it is indistinguishable between one disease or another. In the confusion matrix in the figure, the large number of errors that were generated in the prediction is shown. In view of this problem, the procedure for the multi-class problem was rejected and the binary classification was followed, as shown in the previously mentioned section.

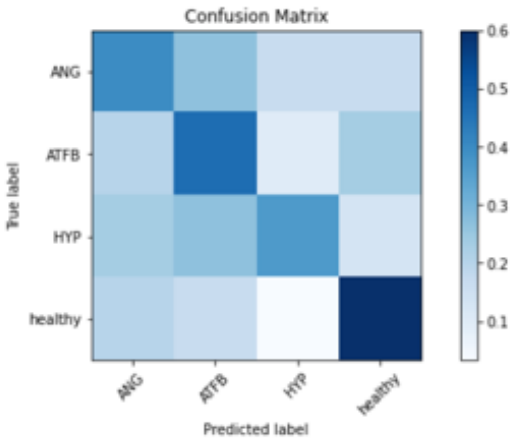

FIGURE B.1: Multiclass UKBB

Appendix C

Feature Selection Alternatives

We considered different methods for feature selection:

- KBest: This method is a filter method, which select features according the K-Highest Score of an statistical Test. For this one we are going to use both *Anova-Test* and *Chi-Suared* Test.
- RFE: This technique correspond to Wrapper Methods and it works by having an external estimator assign coefficients to the features and recursively consider smaller sets of features.
- Sequential Forward Feature Elimination: Automatically select a subset of features that is most relevant to the problem. The goal of feature selection is two-fold: We want to improve the computational efficiency and reduce the generalization error of the model by removing irrelevant features or noise.

For the development of the thesis, we chose KBest for the feature selection technique, since it was less computationally expensive than other two and ultimately, we achieved acceptable results. Nevertheless, as the discussion presented in 4 section 4.4, it generates certain doubts with the features selected in relation to the actual biological component. For that reason, we were interested in performing a few experiments with Sequential Forward Feature Selection (SFFS), due the results obtained by (Cetin et al., 2017).

For this experiment, we reduced the grid search for the number of features and the hyperparameters of the models since one cross-validation could take up a few minutes, leading to extremely large amount of training time for our original grid search. We conducted this experiment only for ACDC dataset and with the best combination of fusion of features 4.5. We set the number of features to 5, 10, 15, 20, 30, 50, 100 and 150. We present the result for the test set using the different number of features. In bold, the best result can be observedC.1. We also provide the feature importance for that specific case C.2. In the results, we can see the potential of this feature selection algorithm, since we reach 0.92 of accuracy. Already higher than our best result. This experiment shows how a more powerful technique for feature selection can improve the accuracy. This technique should be addressed and researched in-depth for future studies.

TABLE C.1: SFS Experiments.

N of features	Accuracy
5	0.8
10	0.9
15	0.88
20	0.88
30	0.88
50	0.9
100	**0.92**
150	0.92

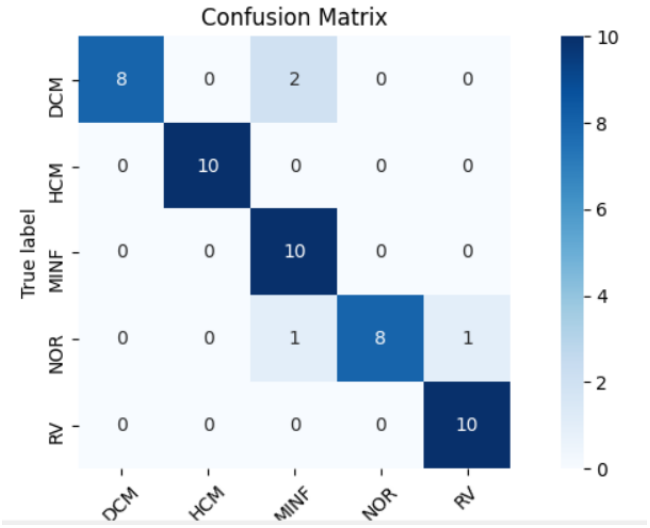

FIGURE C.1: SFS best REsult

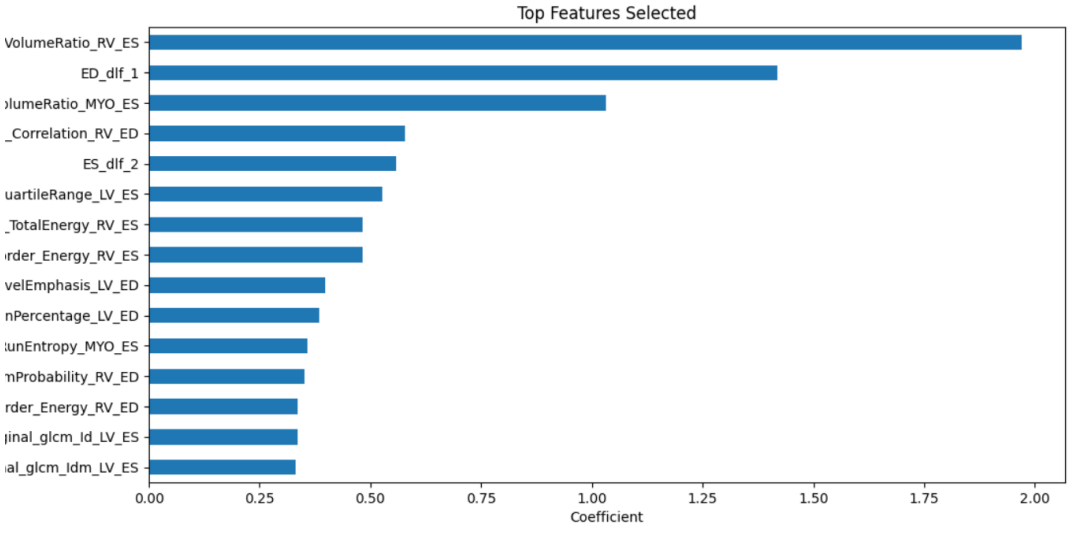

FIGURE C.2: Best Result feature importance

Bibliography

Afshar, Parnian et al. (2018). "From Hand-Crafted to Deep Learning-based Cancer Radiomics: Challenges and Opportunities". In: *CoRR* abs/1808.07954. arXiv: 1808.07954. URL: http://arxiv.org/abs/1808.07954.

Alaa, Ahmed M. et al. (May 15, 2019). "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants". In: *PloS one*. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0213653. published.

Bernard, O. et al. (2018). "Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?" In: *IEEE Transactions on Medical Imaging* 37.11, pp. 2514–2525.

Bizzego, A. et al. (2019). "Integrating deep and radiomics features in cancer bioimaging". In: DOI: 10.1101/568170.

Cetin, Irem et al. (2017). "A Radiomics Approach to Computer-Aided Diagnosis with Cardiac Cine-MRI". In: *ArXiv* abs/1909.11854.

Cui, Ximin et al. (2019). "Multiscale Spatial-Spectral Convolutional Network with Image-Based Framework for Hyperspectral Imagery Classification". In: *Remote Sensing* 11.19, p. 2220. DOI: 10.3390/rs11192220.

Gillies, R., P. Kinahan, and H. Hricak (2016). "Radiomics: Images Are More than Pictures, They Are Data". In: *Radiology* 278, pp. 563 –577.

Griethuysen, Joost J M van et al. (2017). "Computational Radiomics System to Decode the Radiographic Phenotype". In: *Cancer Res* 77.21, e104–e107. ISSN: 1538-7445. DOI: 10.1158/0008-5472.CAN-17-0339.

He, T and et al (Oct. 11, 2019). "Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics". In: *Neuroimage*. URL: https://www.sciencedirect.com/science/article/abs/pii/S1053811919308675?via%3Dihub. published.

Hosny, Ahmed, Hugo J Aerts, and Raymond H Mak (2019). "Handcrafted versus deep learning radiomics for prediction of cancer therapy response". In: *The Lancet Digital Health* 1.3. DOI: 10.1016/s2589-7500(19)30062-7.

Karim, Raimi (2019). *Illustrated: 10 CNN Architecture*. URL: https://towardsdatascience.com/illustrated-10-cnn-architectures-95d78ace614d#bca5.

Khened, Mahendra, Alex Varghese, and G. Krishnamurthi (2019). "Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers". In: *Medical Image Analysis* 51, 21–45.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira et al. Curran Associates, Inc., pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

Lai, Zhifei and Huifang Deng (2018). "Medical Image Classification Based on Deep Features Extracted by Deep Model and Statistic Feature Fusion with Multilayer

Perceptron". In: *Computational Intelligence and Neuroscience* 2018, 1–13. DOI: 10.1155/2018/2061516.

Liu, Siqi et al. (2018). "3D Anisotropic Hybrid Network: Transferring Convolutional Features from 2D Images to 3D Anisotropic Volumes". In: *ArXiv* abs/1711.08580.

Maini, Raman and Himanshu Aggarwal (2010). "A Comprehensive Review of Image Enhancement Techniques". In: *ArXiv* abs/1003.4053.

Martin-Isla, Carlos et al. (2020). "Image-Based Cardiac Diagnosis With Machine Learning: A Review". In: *Frontiers in Cardiovascular Medicine* 7. DOI: 10.3389/fcvm.2020.00001.

Moreira, Catarina et al. (2020). "An Investigation of Interpretability Techniques for Deep Learning in Predictive Process Analytics". In: *ArXiv* abs/2002.09192.

Nie, Dong et al. (2019). "Multi-Channel 3D Deep Feature Learning for Survival Time Prediction of Brain Tumor Patients Using Multi-Modal Neuroimages". In: *Scientific Reports* 9.

Oikonomou, Evangelos K, Musib Siddique, and Charalambos Antoniades (2020). "Artificial intelligence in medical imaging: A radiomic guide to precision phenotyping of cardiovascular disease". In: *Cardiovascular Research*. DOI: 10.1093/cvr/cvaa021.

Patel, Sakshi, P BharathK., and Rajesh Kumar Muthu (2020). "Medical Image Enhancement Using Histogram Processing and Feature Extraction for Cancer Classification". In: *ArXiv* abs/2003.06615.

Raisi-Estabragh, Zahra et al. (Mar. 2020). "Cardiac magnetic resonance radiomics: basic principles and clinical perspectives". In: *European Heart Journal - Cardiovascular Imaging* 21.4, pp. 349–356. ISSN: 2047-2404. DOI: 10.1093/ehjci/jeaa028. eprint: https://academic.oup.com/ehjcimaging/article-pdf/21/4/349/32932013/jeaa028.pdf. URL: https://doi.org/10.1093/ehjci/jeaa028.

Szegedy, Christian et al. (2015). "Rethinking the Inception Architecture for Computer Vision". In: *CoRR* abs/1512.00567. arXiv: 1512.00567. URL: http://arxiv.org/abs/1512.00567.

Wang, Hongkai et al. (2017). "Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18F-FDG PET/CT images". In: *EJNMMI Research* 7.

Woldaregay, Ashenafi Zebene et al. (2019). "Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes". In: *Artificial Intelligence in Medicine* 98, 109–134. DOI: 10.1016/j.artmed.2019.07.007.

Xiang, Yifan et al. (Dec. 2019). "Implementation of artificial intelligence in medicine: Status analysis and development suggestions". In: *Artificial Intelligence in Medicine* 102, p. 101780. DOI: 10.1016/j.artmed.2019.101780.

Yang, H. et al. (2018). "Multimodal MRI-based classification of migraine: using deep learning convolutional neural network". In: *BioMedical Engineering OnLine* 17.