Diagnosis

Daniel Rueckert, Moritz Knolle, Nicolas Duchateau, Reza Razavi, Georgios Kaissis

5 Diagnosis

Prof Daniel Rueckert[a,b,*]

Mr Moritz Knolle[a,b,c]

Dr Nicolas Duchateau[d,e]

Prof Reza Razavi[f]

Dr Georgios Kaissis[a,b,c]

[a] *Artificial Intelligence in Medicine and Healthcare, Technical University of Munich, Munich, Germany.*

[b] *Department of Computing, Imperial College London, London, United Kingdom.*

[c] *Institute of Diagnostic and Interventional Radiology, Technical University of Munich, Munich, Germany.*

[d] *Univ Lyon, Université Claude Bernard Lyon 1, INSA-Lyon, CNRS, Inserm, CREATIS UMR 5220, U1294, F-69621, Lyon, France.*

[e] *Institut Universitaire de France (IUF), France.*

[f] *School of Biomedical Engineering and Imaging Sciences, King's College London, United Kingdom.*

[*] *Corresponding author.*

Authors' contribution:

- Introduction, Opinion: RR.
- Main chapter: DR, MK, GK.
- Tutorial: ND.

Abstract

This chapter covers the clinical application of diagnosis of cardiovascular disease. A clinical opinion piece discusses the current clinical standard for diagnosis tasks and its limitations. The technical review summarizes the classical machine learning pipeline for medical diagnosis as well as some common types of traditional machine learning models that have been used for this application. Following this, some relevant deep learning architectures for computer-aided diagnosis are discussed. Some example applications of artificial intelligence based automated diagnosis are introduced and the key challenges highlighted. The practical tutorial deals with a simple diagnosis task based on characteristics derived from cardiac MR segmentations and other patient characteristics. The chapter closes with a clinical opinion piece that speculates on the future role of AI in cardiac diagnosis.

Keywords:

diagnosis, machine learning, deep learning, artificial intelligence, radiomics, interpretability

Learning Objectives:

At the end of this chapter you should be able to:

- *O5.A Explain the classical machine learning pipeline for medical diagnosis problems*
- *O5.B Describe the key characteristics of commonly used classical machine learning models in diagnosis, such as SVMs and decision trees/forests*
- *O5.C List the types of deep learning architecture that can be applicable to diagnosis problems*
- *O5.D Describe some specific applications for machine learning-based diagnosis in cardiology*
- *O5.E Explain the key challenges involved in the use of machine learning in cardiac diagnosis*

Clinical Introduction

The use of AI in medicine is gaining traction, with many examples moving from research into clinical prototypes and products. Examples include medical record mining [48], predictive clinical decision support systems [299], and, its widest application, the interpretation of medical imaging to help with improving both diagnosis and prognosis of disease [32, 269]. Because of the increasing wealth of digital data that is generated, clinicians need to be able to find more efficient ways of meaningfully combining these data to deliver precision-based medicine. AI can not only enable routine tasks to be performed more efficiently but also provide new insights into disease processes that were previously not achievable by manual review and analysis due to time and labour constraints [91].

Diagnosis and treatment planning of cardiovascular disease is now increasingly reliant on imaging methods such as echocardiography [18], CT [232] and CMR [226]. These generate large amounts of data and yet clinical decision making can often come down to a small number of derived parameters such as the LV EF that use a limited amount of the available acquired imaging information. The application of AI methods to better utilize the available imaging data, overcome challenges with less-than-optimal reproducibility of some of the key biomarkers and reduce the manual workload and time taken to analyse the data is looking promising. AI methods are now being integrated into many clinical products particularly in image analysis [18, 232, 226] but also in image acquisition [232, 226]. Other diagnostic methods such as retinal scanning are also amenable to AI methods. Researchers are now looking to combine the power of AI with the non-invasive ease of retinal scanning to examine the workings of the heart and predict changes in the macrovasculature based on microvascular features and function [148].

In addition to addressing the variability associated with subjective image interpretation, AI can address the spatial and temporal pathologic heterogeneity of cardiovascular clinical phenotypes by allowing more detailed feature extraction around regions of interest [335]. This allows clinicians to use additional quantifiable features that relate more objectively and in more detail to the underlying clinical condition [425]. By extracting a multitude of information generated from images and non-imaging data, AI methods also provide the essential link to uncovering associations between clusters of patients in a fully automated manner [145].

Examples of the use of AI clustering in patients with heart failure have shown that it is possible to identify patient groups with different outcomes, with for example median 21-month survival of 26% vs 63% in patients with heart failure with preserved EF [415], and even different responses to treatment in a larger heart failure cohort [14].

These capabilities will not replace but rather augment the clinical decision process in a more efficient, user-friendly way, that should translate into improved patient care. Recent applications of AI in medical imaging provides proof of concept for its utility and on the whole high performance, with an accuracy paralleling that of human expertise [72, 264].

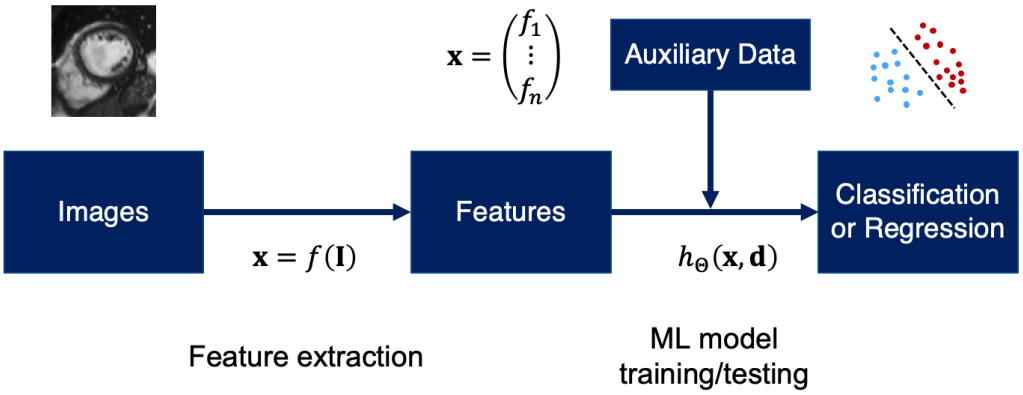


Figure 5.1: Classical machine learning pipeline for diagnosis from cardiac data with the two distinct stages of feature extraction and model learning.

Overview

Over the last decade, AI and machine learning techniques have made significant advances. In particular, as we have seen in Chapter 3 and elsewhere in the book, deep learning [222] has emerged as a powerful framework for solving perceptual tasks across many different application domains, including medicine [386]. Often, deep learning can achieve a level of performance that is comparable to humans (and in some cases even outperforming them) [120, 117]. In this section we will first introduce some machine learning approaches that have been proposed for use in the context of automated medical diagnosis. We begin with classical machine learning approaches before reviewing deep learning approaches. In the subsequent sections we will review their application to diagnosis problems in cardiology as well as discuss challenges for clinical deployment.

Classical Machine Learning Pipeline for Diagnosis

Traditionally, the process of building a system for diagnosis in medicine consists of two stages (see Figure 5.1). In the first stage, information is extracted from the data (e.g. images, signals or clinical data) and in the second stage this information is used to build a statistical model that can perform classification. In machine learning, the information extracted from the data and that is used as input to the statistical model is typically referred to as the

features. In the context of clinical decision making, these features are often referred to as *biomarkers* which serve as measurable indicators of the biological state or condition of the patient. For example, left ventricular myocardial mass or LV EF may be important characteristics when building a diagnosis system for cardiovascular diseases. Additionally, clinical data such as laboratory results (creatine kinase, lactate dehydrogenase, troponine, etc.) or results from other examinations (stress echocardiography, ECG) can be included. In the following we will briefly review some of the most commonly used machine learning models for performing classification using such features. An overview of their performance can be found in Figure 5.2.

Support Vector Machines: The support vector machine (SVM) [94] model, which we first mentioned in Chapter 2, is a very popular algorithm for supervised learning that was first proposed in the early 2000s. It offers robustness and easy applicability to a wide range of problems, domains and types of data without the need for expert prior knowledge. SVMs, which can be used for classification and regression tasks, construct a *maximum margin separator* that defines a decision boundary with maximum distance to its *support vectors*. Specifically, SVMs construct a so-called *soft decision boundary* which is less sensitive to outliers in the data than other approaches. The decision boundary is learned given the training data and assigns classes to data points based on their position in feature space and with respect to the maximum margin separator. This approach incorporates ideas from statistical learning theory [397] to address a common practical problem, namely that for a given dataset (of limited size), there often exist many solutions that split the training data perfectly. For non-linearly separable data, a so-called *kernel function* (see also the Technical Note in Chapter 3, page 72) can be used to transform data points into a higher-dimensional feature space where they become linearly separable (this is often referred to as the *kernel trick*).

SVMs represent a *non-parametric* classification method, meaning that no explicit parameters are learned to define (*parametrize*) the decision boundary. Instead, a set of data points (the support vectors) is used to construct the separating hyperplane in a way that maximizes the distance between the support vectors of the two classes. Of note, the original mathematical formulation for the SVM is only defined for the binary case, however this can be easily extended to the multi-class case by performing *one-against-rest* classification with multiple binary SVMs [96], albeit at a much increased

computational cost. For instance, a binary kernel-trick SVM has a worst-case time complexity of $\mathcal{O}(n^3 \times m)$ (see Technical Note below), where n is the number of training examples and m the number of features. This difficulty of scaling SVMs to large datasets and multi-class prediction as well as the fact that deep neural networks have been shown to outperform SVMs in most applications has led to a drop in their popularity in the more recent machine learning literature. Despite this, SVMs can still be an attractive option for certain use cases (online learning, outlier detection etc.), especially when only a small or intermediate-sized training dataset is available. Furthermore, SVM approaches can also be adapted to regression problems.

Technical Note
The notation $\mathcal{O}(...)$ seen above is known as "Big-O notation". It is commonly used to indicate the *computational complexity* of a task. For example, if n is the number of training samples, $\mathcal{O}(n)$ means that the algorithm takes a time that is proportional to n, $\mathcal{O}(n^2)$ means that the time increases quadratically with n, etc.

Decision Trees and Forests: Decision trees and forests were also mentioned as types of machine learning model in Chapter 2. A decision tree represents a fundamental data structure that can be used to make predictions. While decision trees are used commonly in machine learning, they are also used outside of machine learning, e.g. in operations research, and even in clinical guidelines, to help identify a strategy most likely to reach a goal. In the context of machine learning [97], a decision tree is a tree in which each of the internal (or non-leaf) nodes correspond to a split into sub-trees corresponding to an input feature. Each of the leaf nodes is labelled with a prediction or a probability distribution over multiple predictions. Depending on the type of predictions stored at the leaf nodes, one can differentiate between two types of decision tree. Decision trees where the predicted variable takes continuous values (typically real numbers) are called *regression* trees. Decisions trees where the predicted variable takes categorical values (typically class labels) are called *classification* trees.

Each of the internal nodes corresponds to a split of the training data according to an input feature. Such a split can be thought of as a weak learner since a single split of the training data is unlikely to produce a very accurate

prediction. By creating a set of splits that are hierarchically organized (in the form of a tree) a better prediction can be obtained. Hence, a key step in creating a good set of hierarchical splits is to determine how to split the training data at each node. These splits are typically determined in a top-down fashion by choosing an input feature at each step that "best" splits the training set. Different criteria such as the Gini impurity or information gain can be used to determine the optimal split [331], but in general these approaches aim to measure the homogeneity of the target prediction within the subsets after the split.

While decision trees are a simple and elegant way of building predictors, the performance of decision trees is often limited in real world applications. One way to build stronger predictors is to combine multiple decision trees into so-called decision forests. Such decision forests belong to the class of so-called *ensemble* machine learning methods (see Technical Note, page 95). In these ensemble methods one can differentiate between different approaches in constructing ensembles. In *random forests*, multiple decision trees are built using a technique called *bagging* where the training data are repeatedly resampled using replacement and the final prediction result is obtained by integrating the prediction across different trees using voting schemes. An alternative approach is based on a technique called *boosting* which builds an ensemble classifier by training each new instance to emphasize the training instances that were previously misclassified. The different classifiers are then combined in a weighted voting scheme such as AdaBoost [128].

Deep Learning Approaches for Diagnosis

Deep learning, as introduced in Chapter 3, is based upon the concept of artificial neural networks and offers several advantages for visual information processing, including the ability to learn feature representations with multiple layers of abstraction as well as the ability for end-to-end learning (see Figure 5.3) [222]. Furthermore, deep learning approaches eliminate the need for hand-crafted features and classifiers that otherwise have to be tuned by experts to specific tasks. Instead, it enables end-to-end learning where both the features and the classifiers are directly learned from the available training data. This ensures that the features and classifiers are optimally suited for the task at hand, although this sometimes comes with the cost that the learnt features are not as meangingful (or *interpretable*) to end-users, i.e.

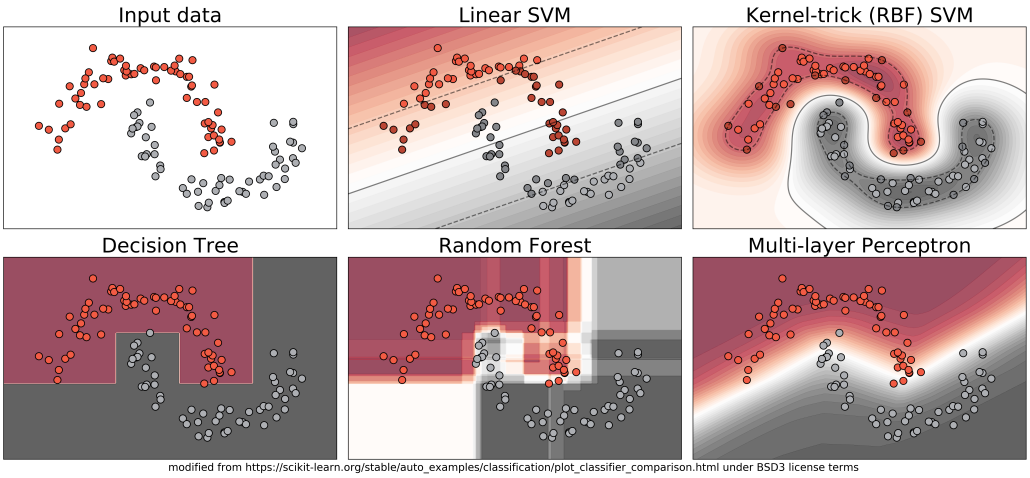


Figure 5.2: Comparison of resulting decision boundaries for different supervised classifiers introduced in this chapter.

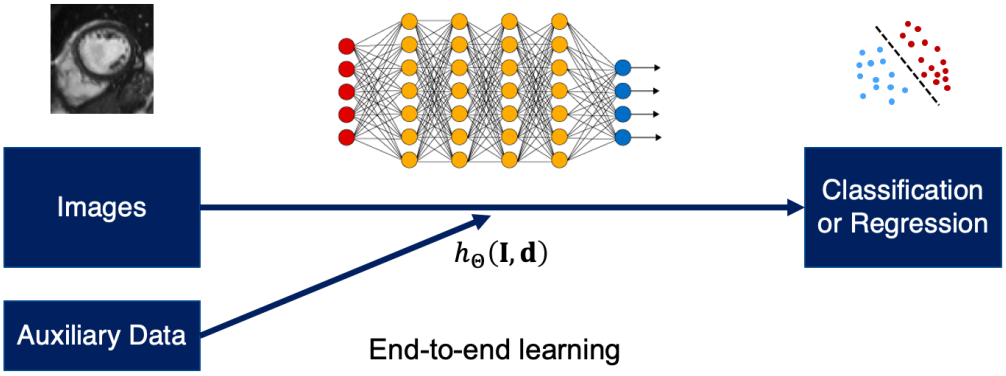


Figure 5.3: Deep learning pipeline with end-to-end trainable feature learning and model learning.

clinicians. Supervised deep learning approaches often employ convolutional neural networks (CNNs) [129, 223] (see Chapter 3, page 74). As we have seen, CNNs consist of many layers that transform their input via convolutions with filters that are learned from the data, making them well suited for images. In contrast, supervised approaches applied to spatio-temporal data (e.g. audio or text) often use recurrent neural networks (RNNs, see Chapter 4, page 88) or long short-term memory (LSTM) networks [165].

In unsupervised deep learning approaches, neural networks based on autoencoders [164] or variational autoencoders [196] are frequently used to reduce the dimensionality of the data (see Chapter 4, Encoder-decoder Networks, page 91). In fact autoencoders also often make use of convolutions in their neural network architecture. Such dimensionality reduction techniques can "simplify" the representation of the data in a way that renders it more conducive to processing by other algorithms, e.g. by subsequent supervised learning algorithms. However, the utilization of unsupervised architectures is not limited to pre-processing, as their output can also be used for diagnosis or scientific discovery. For instance, clustering (see Chapter 2, page 27) can clarify underlying subgroups in datasets, such as groups of patients with similar characteristics, which might, for example, exhibit a common response to a certain medication. Another approach to unsupervised learning is based on generative adversarial networks (GANs, see Chapter 4, page 91) [143] and its variations [31]. As we discussed in Chapter 4, with GANs two neural networks (the generator and discriminator) compete with each other to generate new data with the same statistics as the training set.

In the following sections, we review the most common architectures for deep learning applied to diagnosis problems in more detail[24].

Encoder-decoder networks: Some tasks, such as semantic image segmentation, require *dense* (pixelwise) predictions. Applying CNN architectures to such a problem means that the associated computational complexity scales with the image size. Encoder-decoder networks (EDNs), such as the *fully convolutional network* [241], are a much more efficient approach to addressing dense prediction tasks. Such networks consist of two parts: the *encoder*, in which features are extracted and progressively downsampled, and the *de-*

[24]Editors' note: There is some overlap in content between these descriptions and those provided in Chapter 4 but we choose to include both as we believe they act as complementary perspectives on these important concepts.

coder, in which features (from the encoder) are progressively upsampled to produce an output at the end of the network with identical shape as the input. The rationale behind this technique is the progressive *distillation* of a large number of complex image features in the encoder and their recombination to form new features in the decoder. Label targets to train an EDN are often segmentation masks and hence during training, the network's output is compared to the ground truth label map via overlap measures such as the commonly used Dice coefficient [374]. Parameter updates are then applied iteratively using a gradient-based optimization method such as stochastic gradient descent (see Chapter 3, page 61). A fully convolutional network is such an EDN where only convolutional and pooling layers (see Technical Note, page 107) are used to extract and process features. A commonly used EDN architecture for medical image segmentation is the *U-net* [332], where skip-connections from encoder to decoder were added to the fully convolutional network architecture, improving the convergence, performance and robustness of the original architecture.

Generative adversarial networks (GANs): *Generative adversarial networks* (GANs) [143] are a relatively recent neural network architecture and training paradigm, whereby two competing neural networks (generator and discriminator) are trained simultaneously to produce a powerful generative model. More precisely, the *generator*, conditioned on random noise vectors, generates artificial data samples while the *discriminator* tries to determine whether they are fake or belong to the target population. While in theory, GANs are capable of approximating any data-generating distribution given enough training data, the training process of GANs is often unstable and sensitive to the choice of hyperparameters. This can be caused by a multitude of reasons, but most of them relate to a disparity in the learning progress of the generator and the discriminator. As a result, countless variations and improvements of the original implementation have been proposed to tackle these problems (an overview and applications for medical imaging can be found in [428]).

Conditional GANs (cGANs), such as the Pix2Pix GAN [172] are a particularly interesting GAN (re)-formulation, where the generator can be conditioned on additional input information (e.g. images). This more sophisticated sampling method allows them to transfer features from one image to the other. Conditional GANs can be used for so-called *domain adaptation*, whereby, for example, CT images can be transformed into *virtual MR im-*

ages. Another potential medical use lies in the generation of training data in scenarios where data are scarce [252] and/or there are associated privacy concerns, as GANs can be used to generate realistic looking, yet fake and thus private medical image data [387]. Additionally, GANs can be used for classification tasks by using part of either the generator or discriminator as a feature extractor, or alternatively by using the discriminator as a classifier. These GAN-based classification approaches have been shown to perform on-par with supervised neural network architectures, but require much less data while also potentially limiting the effect of domain overfitting [253].

Autoencoders: As mentioned earlier, an autoencoder [164] is a type of EDN that transforms (often high dimensional) input data into a lower dimensional *latent vector representation* in an unsupervised fashion. An autoencoder learns the optimal latent representation of the training data by attempting to reconstruct the original input data solely based on the encoded latent information. While autoencoders are an excellent dimensionality reduction technique, the resultant latent space representation is often incomplete and not optimally suited for generative sampling purposes. *Variational autoencoders* [196] explicitly learn to parametrize a Gaussian distribution from which sampling is performed. This more principled approach makes VAEs much better suited for generative tasks than conventional autoencoders. However, the images produced by VAEs usually look much less realistic than GAN-produced images, especially when high resolution images are desired.

Bayesian deep learning: The *Bayesian* inference framework, based around Bayes' theorem (the principled approach that prior assumptions influence posterior beliefs), offers the most complete approach for reasoning under uncertainty and is therefore a key component for building real-world systems in safety-critical applications such as self-driving cars or computer-aided diagnosis tools. Quantifying predictive confidence as well as uncertainty-based human expert referral [225] are thus crucial for establishing and promoting trust in an automated diagnosis system when deployed in a real-world clinical setting. In practice, traditional (point-estimate) neural networks are often over-confident about their predictions, thus highlighting the need for a sound approach to modelling uncertainty in deep learning. In Bayesian neural networks (BNNs) [8] each parameter is represented using a (posterior) probability distribution to model uncertainty. The computation of this distribution is, however, usually intractable due to the requirement to calculate

high dimensional integrals for its precise specification.

Two main techniques are used to avoid this intractable computation: *variational inference* and *Markov Chain Monte Carlo* (MCMC) [158]. In variational inference, the posterior distribution is approximated by a simpler (*variational*) distribution which is learned during training by optimising the *distributional similarity* between the variational distribution and the true posterior. On the other hand, MCMC methods take a sampling-based approach to computing the posterior by randomly drawing *samples* from areas of the posterior with high probability density. Each of the methods has its own benefits and drawbacks. In general, it can be stated that variational inference tends to be significantly faster, however it produces a biased estimate of the posterior distribution. MCMC is capable of exactly representing the posterior. However, it is both slower and computationally more expensive. In recent years, these methods have been complemented by newly-proposed *approximate* methods, empirically shown to enable reasonable uncertainty estimates without requiring the utilization of the above-mentioned inference techniques. For example, Monte Carlo dropoutDropout [131] utilizes a technique originally proposed for regularization and DeepEnsembles [206] leverages ensembles of neural networks to quantify predictive uncertainty. These two methods are performant and easy-to-implement approaches to making any neural network *Bayesian*.

Machine Learning Applications for Diagnosis

As described above, the computer-aided diagnosis of cardiovascular diseases plays an increasingly important role in clinical routine. A task that is commonly addressed using machine learning approaches is the classification of different cardiac pathologies. For example, the Automatic Cardiac Diagnosis Challenge (ACDC) [50], which primarily focuses on cardiac image segmentation, also proposes to diagnose different diseases with abnormal myocardial shape, including in addition to normal subjects, (1) patients with systolic heart failure with infarction, (2) patients with dilated cardiomyopathy, (3) patients with hypertrophic cardiomyopathy and (4) patients with abnormal RV. Many machine learning approaches use this dataset as benchmark for cardiac disease classification [170, 194, 68, 412].

The majority of recent approaches use deep learning for disease classification, using information about cardiac morphology as well as cardiac function.

However, these approaches often do not allow for easy interpretation of the classification results. In [53], the authors tackle this problem by developing an interpretable deep learning model for disease classification using cardiac shape information. They exploit deep generative networks to model a population of anatomical shapes through a hierarchy of conditional latent variables. The approach has been shown to provide high classification accuracy as well as visualization of both global and regional anatomical features which discriminate between different pathologies. The interpretability of deep learning approaches is also the focus of the work in [89], in which a CNN model is used together with a VAE to learn a discriminative latent space for classification. Using the idea of 'concept activation vectors' [195], the latent space is then visualized in terms of diagnostically meaningful clinical parameters.

In [441] the authors classify different cardiac pathologies by combining features derived from segmentations of the cardiac anatomy, their shapes and motion patterns. A similar approach is pursued in [316] which uses a multimodal database of CMR and echocardiography images to learn cardiac motion patterns. During inference only motion from the echocardiography images is used to discriminate between normal subjects and patients with dilated cardiomyopathy. Other approaches that focus on the analysis of cardiac function from echocardiography images [292] have gained a lot of attention due to the wide availability of this modality. However, not all approaches focus on using image data as the primary source of information. For example, ECG data is a widely available source of important physiological information about cardiac abnormalities and analysis of ECG signals using machine learning approaches can provide powerful diagnostic tools [22].

Machine Learning Approaches Based on Radiomics

In the context of cardiovascular imaging so-called radiomics approaches also play an important role for diagnosis. Radiomics approaches aim to extract a large number of shape- or texture-based features from images that may then be used as predictor variables in statistical models for diagnosis. Radiomics has been successfully used in oncology [13] and more recently also in cardiology [68, 324]. The success of radiomics approaches depends heavily on the type of images used as the reproducibility of the extracted shape and texture information is critical for the success and reliability of these approaches. Furthermore, standardization of the imaging data is crucial when data from

multiple hospitals or imaging centres is used.

Machine Learning Approaches for Large-Scale Population Studies

Machine learning approaches also play an important role in discovering quantitative and clinically relevant phenotypes from population studies, which can in turn promote the discovery of novel diagnostic biomarkers. In [37], the authors used a deep learning pipeline for extracting 82 quantitative phenotypes of the heart and aorta from CMR from a large population study with over 25,000 participants from the UK Biobank [307]. They identified 2,617 significant associations between imaging phenotypes and non-imaging phenotypes of the participants describing relationships between risk factors and cardiovascular diseases.

While the large-scale extraction of biomarkers and phenotypes from population studies is challenging, it is also important to perform quality control of the information extracted from such studies. For example, in CMR studies, the extraction of biomarkers may fail because of poor image quality or image artefacts (e.g. respiratory motion) or the image analysis pipeline may fail, and therefore affect the downstream task such as diagnosis. To address this problem, it is possible to use machine learning techniques to classify whether the image quality is sufficient for automated analysis [381] or whether the extracted parameters are likely to be correct [335, 382]. We return to the topic of automated quality control in cardiac image analysis in Chapter 7.

Challenges

Despite the significant advances in the development of machine learning approaches in cardiology, there remain a number of challenges. One of the challenges is that deep learning approaches tend to require significant amounts of training data. In general, the more data are available for training, the more accurate and robust the resulting machine learning models become. The need for large datasets and high quality annotations makes data sharing even more important, not only for training but also for evaluating machine learning solutions in multi-institutional/multi-national trials. One solution to this challenge has been found in the availability of large datasets from prospective volunteer trials (such as the UK Biobank [373]) or from curated

clinical databases such as PhysioNET [142]. However, in practice data sharing is often hampered by technical, legal and ethical challenges. In particular, legal and regulatory requirements represent difficulties for data sharing. An alternative to data sharing is the use of decentralized machine learning or federated learning approaches [329, 184]. In contrast to centralized approaches in which datasets are marshalled in one central location to train one machine learning model, federated learning uses collaborative training algorithms that do not require the exchange of the training datasets with a central instance. It has been shown that these federated learning approaches can achieve similar performance to conventional centralized approaches and outperform approaches that are only trained using data from one site.

Another challenge for clinical adoption of machine learning-based approaches is the perceived black box nature of many of these approaches. This means that the output of a diagnosis by a machine learning model can be difficult for humans to understand and interpret. Recent guidelines of the European Union emphasize the importance of explainability and interpretability of AI-based approaches, especially if they affect humans directly. However, there is a lack of consensus as to precisely what explainability and interpretability mean in this context. Related to this challenge is also the fairness of decision-making algorithms. Fairness can be defined as the absence of any prejudice or bias toward an individual or a group based on a set of protected characteristics such as race, sex or age. It can be difficult to detect biases and unfairness in machine learning approaches that learn from data. The source of such problems is often (but not always) related to biases and/or imbalance in the data that are used to train the machine learning models. Identifying these biases is a first step to mitigating for the bias and developing "fair" machine learning approaches.

Closing Remarks

Whilst techniques for automated machine learning-based diagnosis in cardiac imaging are less mature than those for measurement and quantification, significant progress has been made in recent years, partly due to the availability of public databases for some cardiac diagnostic tasks. As in most applications, deep learning models are currently the best performing techniques in diagnosis in terms of classification accuracy, although classical machine learning models likely still have a role to play, especially in less complex

tasks with a limited amount of annotated data. Deep learning models are seen as being less interpretable than some classical machine learning models, but researchers have taken note of the need for interpretability in diagnostic tools and have proposed methodological advances to address this issue. These interpretability techniques need further evaluation on real-world clinical data, and their role in clinical workflows needs to be carefully considered and their impact well validated. Of particular concern is the possibility for bias, for example based on the sex or race of the subject: a recent work [319] has shown the potential for racial bias in diagnosis of heart failure based on deep learning-derived LV EF measurements made from CMR imaging. Open and complete reporting of performance across such subgroups is therefore of paramount importance [283]. In the following tutorial you will gain practical experience of using deep learning for a simple diagnostic task.

Exercises

Exercise 1.
Explain the main difference between classical machine learning and deep learning approaches with regard to the features used for automated diagnosis.

Exercise 2.
Explain what is meant by "end-to-end learning" in the context of deep learning-based diagnosis.

Exercise 3.
A colleague argues that machine learning-based diagnosis will never be completely trusted by cardiologists. Therefore, we should consider their use as decision support tools rather than automated diagnosis tools. Do you agree? What implications would this have for the design of such tools?

Exercise 4.
What role do you see Bayesian deep learning playing in automated diagnosis?

Exercise 5.
A colleague argues that dealing with possible bias in machine learning-based diagnosis is less important than optimizing overall performance. Do you agree?

Exercise 6.
A research group is working with cardiologists to develop a tool for automated diagnosis of some rare cardiovascular diseases from cine CMR imaging. Advise the group on what type(s) of machine learning approach might be applicable and what issues they should be aware of.

Exercise 7.
Supervised machine learning might seem the preferred approach for automated diagnosis, since the task is to predict a label (diagnosis) given the available data. What role could unsupervised machine learning have in automated diagnosis?

Tutorial - Two-class and Multi-class Diagnosis

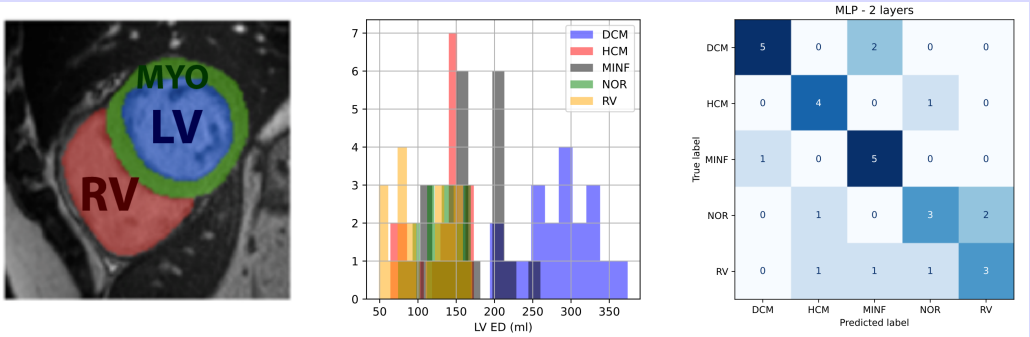
Tutorial 4.

As for the other notebooks, the contents of this notebook are accessible as Electronic Supplementary Material.

Overview

In this hands-on tutorial, you will again use data from the ACDC open access dataset [1]. this time not for segmentation but for diagnosis, based on characteristics extracted from the image segmentations and additional patient characteristics. This corresponds to the task targeted in the second part of the paper reporting on the ACDC challenge [50].

You will use data from the 100 patients of the training set, which are equally distributed into 5 (ab)normal subgroups. The tutorial will guide you through the classification of these subjects, starting from two-class diagnosis (e.g. normal vs. dilated hearts) and moving to the more complete multi-class diagnosis, as illustrated in the figure below. The tutorial lays stress on carefully examining the performance against the complexity of the machine learning model, keeping in mind the data used as input.



Descriptors extracted from MR images

Classification output

Objectives

- *Consolidate the knowledge you've gained on classification from the toy examples in the hands-on tutorial from Chapter 3.*
- *Conduct a proper classification problem on real-life data.*
- *Get used to a wider variety of* `scikit-learn` *models and be critical about their ouptut.*

Computing Requirements

As for the other hands-on tutorials, this notebook starts with a brief "System setting" section, which imports the necessary packages, installs the potentially missing ones, and imports our own modules.

Opinion

There are ethical, regulatory and practical challenges that need to be addressed to ensure reliability, quality of care and safety before wide scale adoption into clinical prime time [106, 121, 308]. A further challenge is the difficulty in replicating the performance of often complex models built on local databases in other clinical settings. The use of federated learning techniques and infrastructure to build models with much wider and varied data sets across multiple clinical settings and geographies could be a good way of addressing this [329, 184] (see also Chapter 10, page 246). The infrastructure to easily deploy models into a clinical setting for different clinical service delivery organizations with different IT systems can also be a challenge that needs to be addressed. Finally, the health economic case will need to be made for individual applications, alongside their clinical utility, as pressure on healthcare budgets will otherwise make commercial success and wide procurement of diagnostic and decision support systems that use AI difficult.

Nevertheless, the application of AI has the potential for reproducible clinical assessments through automated measurements, more efficient diagnostic support, improved phenotyping and better risk stratification through the mining of large datasets to uncover clinically relevant information [347]. Its application to care of patients with cardiovascular disease will be transformative and bring substantial benefit.

Acknowledgements

ND was supported by the French ANR (LABEX PRIMES of Univ. Lyon [ANR-11-LABX-0063] within the program "Investissements d'Avenir" [ANR-11-IDEX-0007], and the JCJC project "MIC-MAC" [ANR-19-CE45-0005]).