UNIVERSITÀ DEGLI STUDI DI TRIESTE

XXXVII CICLO DEL DOTTORATO DI RICERCA IN

APPLIED DATA SCIENCE AND ARTIFICIAL INTELLIGENCE

MACHINE LEARNING APPLICATIONS IN CARDIOLOGY

Settore scientifico-disciplinare: INF/01

DOTTORANDO
GIOVANNI BAJ

COORDINATORE
PROF. FRANCESCO PAULI

SUPERVISORI DI TESI
PROF. GIULIA BARBATI
PROF. LUCA BORTOLUSSI
DR. ARJUNA SCAGNETTO

ANNO ACCADEMICO 2023/2024

UNIVERSITY OF TRIESTE

DEPARTMENT OF MATHEMATICS, INFORMATICS AND GEOSCIENCES

PHD IN APPLIED DATA SCIENCE AND ARTIFICIAL INTELLIGENCE

MACHINE LEARNING APPLICATIONS IN CARDIOLOGY

GIOVANNI BAJ

Matr. PHD1400007

THESIS SUPERVISORS:

Prof. Giulia Barbati
Prof. Luca Bortolussi
Arjuna Scagnetto

ACADEMIC YEAR 2023-2024

ABSTRACT

Cardiovascular diseases remain the leading cause of death globally and impose significant economic burdens, emphasizing the need for advanced prevention and management strategies. This thesis investigates the transformative potential of artificial intelligence in cardiology, employing state-of-the-art machine learning and deep learning methodologies to address key challenges in cardiovascular care.

Using electrocardiogram data, the first part of this work evaluates machine learning models for atrial fibrillation prediction, demonstrating the impact of class imbalance corrections on calibration and the importance of adequate sample sizes for deep learning, explored through learning curve techniques. Continuing within the context of atrial fibrillation prediction, multi-modal approaches integrating electrocardiogram and tabular data in survival frameworks are compared to single-modality models. Unsupervised clustering is applied to phenotype dilated cardiomyopathy patients using electrocardiogram, demographic, and clinical data, identifying clusters with distinct genetic backgrounds and outcomes. Additionally, a novel method that combines survival neural networks with functional clustering is developed to assess treatment response, capturing time-dependent effects and feature interactions for personalized care. In the extended framework of integrating also images, a fully automated deep learning tool for coronary computed tomography angiography was developed, demonstrating high accuracy in detecting rare congenital heart diseases. Finally, machine learning-based diagnostic models effectively identified transthyretin amyloid cardiomyopathy in patients with severe aortic stenosis, with computed tomography strain emerging as the most accurate modality.

The findings of this thesis highlight the potential of artificial intelligence tools to enhance cardiovascular diagnostics and patient management, emphasizing the critical role of rigorous study design and the careful evaluation of performance measures aligned with clinical objectives

PUBLICATIONS

PUBLISHED

- **Giovanni Baj** et al. "Comparison of discrimination and calibration performance of ECG-based machine learning models for prediction of new-onset atrial fibrillation." In: *BMC Medical Research Methodology* 23.1 (July 2023), p. 169. DOI: 10.1186/s12874023-01989-3
- **Giovanni Baj** et al. "Deep Learning Survival Model to Predict Atrial Fibrillation From ECGs and EHR Data." In: *Progress in Artificial Intelligence*. Springer Nature Switzerland, 2023, pp. 222–233. DOI: 10.1007/978-3-031-49011-8_18
- Isaac Shiri, Sebastian Balzer, **Giovanni Baj** et al. "Multi-modality artificial intelligence-based transthyretin amyloid cardiomyopathy detection in patients with severe aortic stenosis." In: *European Journal of Nuclear Medicine and Molecular Imaging* (Sept. 2024). DOI: 10.1007/s00259-024-06922-4

UNDER REVIEW

- Ilaria Gandin, Maria Perotto, Alessia Paldino, **Giovanni Baj** et al. "Clustering in Dilated Cardiomyopathy at Initial Evaluation: An Effective Tool for Clinical Stratification". Submitted to *European Journal of Heart Failure*.
- Isaac Shiri, **Giovanni Baj** et al. "Artificial Intelligence Based Detection and Classification of Anomalous Aortic Origin of Coronary Arteries in Coronary CT Angiography: A Multi-Center Development, Testing and Clinical Evaluation Study". Under review in *Nature Communications*.
- Daniela Pacella, Emanuele Giusti, Annamaria Porreca, **Giovanni Baj**, Ilaria Gandin, Giulia Barbati. "Sample size determination via learning curves for AI models: an application to deep learning algorithms for diagnostic and prediction tasks". Submitted to *Artificial Intelligence in Medicine*.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AAOCA	Anomalous Aortic Origin of Coronary Arteries
AF	Atrial Fibrillation
AI	Artificial Intelligence
AS	Aortic Stenosis
ASUGI	Azienda Sanitaria Universitaria Giuliano Isontina
AUC	Area Under the Receiver Operating Characteristic Curve
ATTR-CM	Transthyretin Amyloid Cardiomyopathy
CAD	Coronary Artery Disease
CATE	Conditional Average Treatment Effect
CCTA	Coronary Computed Tomography Angiography
CECT	Contrast-Enhanced Computed Tomography
CHARGE-AF	Cohorts for Aging Research and Genomic Epidemiology
CI	Confidence Interval
CIF	Cumulative Incidence Function
CMR	Cardiovascular Magnetic Resonance
CNN	Convolutional Neural Network
CRT	Cardiac Resynchronization Therapy
CT	Computed Tomography
CTA	Computed Tomography Angiography
CV	Cross-validation
CVD	Cardiovascular Disease
DCM	Dilated Cardiomyopathy
DL	Deep Learning
DNN	Deep Neural Network
ECG	Electrocardiogram
EHR	Electronic Health Records
ESC	European Society of Cardiology
FCN	Fully Connected Network

FDA	Functional Data Analysis
FVG	Friuli Venezia Giulia
GBD	Global Burden of Disease
GP	Gaussian Process
HF	Heart Failure
HT	Heart Transplantation
ICI	Integrated Calibration Index
IQR	Interquartile Range
KM	Kaplan-Meier
L-AAOCA	Left Anomalous Aortic Origin of Coronary Arteries
LASSO	Least Absolute Shrinkage and Selection Operator
LA-GLS	Left Atrial Global Longitudinal Strain
LBBB	Left Bundle Branch Block
LR	Logistic Regression
LV	Left Ventricle
LV-GLS	Lef Ventricle Global Longitudinal Strain
LVEF	Left Ventricular Ejection Fraction
LVH	Left Ventricular Hypertrophy
LVMi	Left Ventricle Mass index
ML	Machine Learning
MLP	Multi-layer Perceptron
MRMR	Minimum Redundancy Maximum Relevance
MVA	Malignant Ventricular Arrhythmias
NDLVC	Non-Dilated Left Ventricular Cardiomyopathy
NN	Neural Network
NLS	Non-linear Least Squares
NSVT	Non-Sustained Ventricular Tachycardia
NYHA	New York Heart Association
PC	Principal Component
PCA	Principal Components Analysis
PCAmix	Principal Components Analysis for mixed data
PH	Proportional Hazards
PMF	Probability Mass Function
PVC	Premature Ventricular Contraction
R-AAOCA	Right Anomalous Aortic Origin of Coronary Arteries
ReLU	Rectified Linear Unit
RER	Regional Epidemiological Repository
RF	Random Forest
RFE	Recursive Feature Elimination
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristic

R-PSM	Royston-Parmar Survival Model
RUS	Random Under Sampling
RV	Right Ventricle
SE-resNet	Squeeze-and-Excitation residual Network
SCD	Sudden Cardiac Death
SD	Standard Deviation
SE	Standard Error
SGD	Stochastic Gradient Descent
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Oversampling Technique
SPECT	Single-Photon Emission Computed Tomography
SSD	Sample Size Determination
SVM	Support Vector Machines
t-SNE	t-distributed Stochastic Neighbor Embedding
TAVI	Transcatheter Aortic Valve Implantation
VF	Ventricular Fibrillation
XGB	eXtreme Gradient Boosting

INTRODUCTION

AI IN CARDIOLOGY

Cardiovascular Diseases (CVDs), which encompass a range of heart and blood vessel disorders, are the leading cause of death globally and a significant contributor to disability [1]. With the rise in population and aging demographics, the prevalence of CVD is projected to increase markedly. [2]. According to the Global Burden of Disease (GBD) Study, a multinational research study that estimates disease burden for every country in the world, the number of CVD cases doubled from 271 million cases in 1990 to 523 million in 2019 [2]. Similarly, CVD-related deaths rose from 12.1 million to 18.6 million over the same period [2]. The economic impact of CVDs is also substantial. In the United States, costs associated with CVD reached an estimated $555 billion in 2016, with projections expecting this number to rise to $1.1 trillion by 2035 [3]. This estimate includes direct medical expenses such as hospital care and medications, as well as indirect costs like lost productivity due to illness, premature death, or informal caregiving. The trend is similar in Europe, where CVD accounted for €282 billion in costs in 2021, with health and long-term care expenses totaling €155 billion (representing 11% of Europe's health expenditure) [4].

These statistics highlight the urgent need for enhanced prevention and management strategies to address the growing burden of cardiovascular diseases. Artificial Intelligence (AI) holds great potential for improving cardiology care across various tasks, including increasing diagnosis accuracy, personalized treatments, enhancing image quality, streamlining clinical workflows, and facilitating the discovery of new digital biomarkers for risk assessment. The recent rise in AI popularity has been propelled by the advancements in Machine Learning (ML). Unlike conventional rule-based symbolic AI, ML involves computer algorithms that learn patterns directly from data [5]. Deep Learning (DL), a subset of ML based on deep artificial neural networks, plays a fundamental role by enabling the analysis of complex data structures like images and signals. The proliferation of AI applications in healthcare is evident from the number of AI-based algorithms approved by the U.S. Food and Drug Administration, which exceeds 900 since 1995. Notably, more than half of these algorithms have received approval in the last 2 years [6]. Of these AI-approved algorithms, 10% are specifically dedicated to cardiovascular applications.

One area where AI has shown remarkable value is the analysis of the Electrocardiograms (ECGs) [7], especially in arrhythmia detec-

tion and classification, achieving diagnostic performance comparable to experienced physicians [8–10]. AI has also allowed the accurate prediction of paroxysmal Atrial Fibrillation (AF) from a 12-lead ECG recorded in normal sinus rhythm [11, 12], probably reflecting the ability to detect modest atrial perturbations that human readers might miss. AI-enhanced ECG interpretation has expanded to include diagnoses of left ventricular dysfunction [13, 14], valvular heart disease [15], channelopathies [16], and even systemic conditions like hyperkalemia [17] and anemia [18]. In echocardiography, AI has been applied to advance both image acquisition and interpretation. AI platforms are being developed to assist novice users in obtaining standard echocardiography views, potentially enhancing point-of-care diagnosis, especially in primary care and emergency settings [19]. Automated tools have been proposed to automate image segmentation, reducing the time needed for clinicians to measure quantities such as left ventricular ejection fraction, and also minimizing variability in assessments [20]. AI is also advancing other cardiac imaging techniques, such as Cardiovascular Magnetic Resonance (CMR) and cardiac Computed Tomography (CT). In CMR, AI improves image quality and speeds up acquisition using denoising techniques for image reconstruction [21], and it automates segmentation of the ventricles and atria in cine images [22]. In cardiac CT, it has been shown that AI can quantify coronary calcium scores [23], assess stenosis and plaque burden [24], and even identify coronary calcifications on non-gated scans [25]. As for CMR, AI has been integrated into all major vendors' tools for image reconstruction, automatic segmentation, and motion correction [7]. AI-based risk prediction models have been developed with both CMR and CT images, for example to predict general survival in patients with pulmonary hypertension [26] and to predict major adverse cardiovascular events in patients with suspected coronary artery disease [27].

Despite the rapid increase in studies focusing on AI technologies for cardiovascular care, these algorithms are still in the early stages of clinical integration, and real-world applications have yet to meet expectations fully [7]. Reasons for this may include training data quality, lack of models' interpretability, and the absence of external evaluation [28]. For instance, among the devices approved by the U.S. Food and Drug Administration up to 2022, only 56% of them reported a clinical validation [29]. To make these tools applicable in real-world scenarios it is fundamental to adopt rigorous methodologies in model development, from study design to evaluation [30]. In this context, the first part of this thesis focuses on methodological aspects of clinical prediction models tackling crucial issues such as evaluating the performance of algorithms in terms of discrimination and calibration and determining the optimal sample size for a study. Another possible weakness of current AI models is that they are built to look at data

from 1 single modality, while CVD is heterogeneous and requires the integration of multimodal data to understand and address complex underlying mechanisms of disease [31]. This topic is also explored in this thesis work in the context of atrial fibrillation prediction. The remaining part of the thesis focuses on specific cardiovascular problems, applying both unsupervised and supervised ML techniques.

THESIS OUTLINE

Chapter 1 provides an introduction to machine learning and deep learning, with a particular emphasis on the network architectures employed in the studies discussed. It also outlines the diseases and the related research questions explored in the remaining chapters, alongside the types of data used to develop the models.

Chapter 2 focuses on risk prediction models for atrial fibrillation using ECG data. The first part examines the discrimination and calibration performance of two machine learning models when corrections for class imbalance are applied in model training. The second section focuses on sample size determination for deep learning models using a learning-curve approach, with the prediction and detection of AF serving as case studies. Finally, the last part of the chapter investigates the integration of ECG and tabular data for AF prediction within a survival framework, accounting for death as a competing risk. The multi-modal survival approach is then compared to single-modality and binary classification methods.

Chapter 3 explores unsupervised clustering in two different scenarios. In the first study, dilated cardiomyopathy patients are phenotyped into different clusters using only information available at first medical contact (ECG, demographic and clinical data). Then these clusters are evaluated in terms of genetic background and outcomes. In the second study, a novel method to assess treatment response in a survival setting is presented.

Chapter 4 presents a study that develops a ML model to detect a rare congenital heart disease (anomalous aortic origin of the coronary arteries) using 3D coronary computed tomography angiography images. The fully automated tool includes coronary centerline detection, coronary segmentation, detection, and risk classification. The model is then tested on an external dataset.

Chapter 5 focuses on developing and evaluating machine learning models using pre-procedural and routine data from Transcatheter Aortic Valve Implantation procedures to detect Transthyretin Amyloid Cardiomyopathy. It compares the performance of different diagnostic modalities, including echocardiography and CT strain, in predicting Transthyretin Amyloid Cardiomyopathy (ATTR-CM) in patients with severe Aortic Stenosis.

Finally, the conclusion summarizes the main contributions and limitations of this work, while discussing the possible future directions of the methods presented.

1 BACKGROUND

1.1 INTRODUCTION TO ML AND DL

The birth of AI can be traced back to 1956, when a group of researchers led by John McCarthy organized a two-month workshop specifically aimed at exploring how to "make machines use language, form abstractions, and concepts, solve kinds of problems now reserved for humans, and improve themselves" [32]. Since then, many definitions of AI emerged. Broadly speaking, AI can be defined as a branch of computer science focused on creating intelligent machines capable of performing tasks that typically require human intelligence, such as speech and pattern recognition, problem-solving, and learning from experience. Today, the term AI generally refers to software and technologies based on machine learning and deep learning, two rapidly advancing fields that have seen significant growth in recent years.

1.1.1 *Machine Learning*

ML is a branch of artificial intelligence that lies at the intersection of several fields, including computer science, statistics, mathematics, and neuroscience. Differently to previous artificial intelligence approaches that relied on hard-coded knowledge [33], ML focuses on developing algorithms and statistical models that enable computers to learn and improve their performance from experience without being explicitly programmed [34]. In computer systems, experience is represented by data, and the primary goal of ML is to create algorithms that can generate models from data in order to make predictions on new observations. Given a dataset, models are usually generated using only a subset of the entire dataset, called *training set*. The rest of the data is named *test set*, and it is used to verify the *generalization* ability of the model, i.e. to check if the model performs well on unseen samples.

Based on the learning paradigm, ML can be divided into different classes. These include [35]:

- *Supervised learning*: The dataset is labeled, i.e. each example in the dataset has an associated label, also known as *outcome*. The ML model is trained to predict the label for a given sample. The label can be discrete (classification), or continuous (regression).

- *Unsupervised learning*: The dataset is unlabelled. In this case, the goal is to discover underlying structures and patterns from unlabeled training samples. Key applications of unsupervised learning include clustering, dimensionality reduction, and density estimation. Specifically, clustering aims to divide the dataset into different disjoint subsets, called *clusters*.

- *Semi-supervised learning*: The training set is only partially labelled. Both labeled and unlabeled data are used to train the model. This is a typical setting where labels are difficult or expensive to obtain, like the medical field [36].

- *Active learning*: The dataset is only partially labeled, and the algorithm interactively queries a human expert to label new data points. The goal is to achieve performance comparable to standard supervised learning but with fewer labeled instances. As for semi-supervised learning, this approach finds application in medicine when labels are scarce [37].

- *Reinforcement learning*: The algorithm learns to make decisions by interacting with an environment and receiving feedback through rewards and penalties.

- *Self-supervised learning*: It involves training models on unlabeled data through pretext tasks to learn underlying data structures. The model is then fine-tuned on a smaller labeled dataset to perform a specific task [38].

- *Transfer learning*: With this technique, the knowledge learned from a task is re-used to boost performance on a related task.

Machine learning encompasses a wide set of models, including among the others decision trees [39], support vector machines [40], artificial neural networks [41], and k-Nearest Neighbors [42].

A popular technique in ML is ensemble learning [43], in which multiple simple models are trained to solve a certain learning problem, and then their predictions are combined to get a single common prediction. It has been shown that this approach can lead to better generalization ability compared to single models [34]. The two most typical ensemble methods are *boosting* and *bagging*. The boosting method trains the base learners (i.e. the individual models) sequentially and with strong correlations, in such a way that incorrectly classified samples in the training set receive more attention from subsequent base learners. A sub-family of boosting algorithms is *gradient boosting machines*, in which the base learners are constructed to be maximally correlated with the negative gradient of the loss function [44]. A famous tree gradient boosting framework that is scalable and that has been used by the data science community to solve a large number

of problems is eXtreme Gradient Boosting (XGB), also known as XGBoost [45]. The second ensembling method, bagging, trains the base learners in parallel and it's based on bootstrap sampling [46]. Random Forest algorithm is an extension of bagging, that uses decision trees as base learners [47].

1.1.2 *Deep Learning*

DL is a branch of machine learning that utilizes neural networks composed of multiple layers, which gives it the term "deep." DL has revolutionized the field of ML, significantly advancing the state-of-the-art in various tasks, particularly those involving the analysis of complex data such as images, videos, text, speech, and audio. The recent rapid growth of deep learning has been possible thanks to the availability of large, high-quality datasets and to the advancements in hardware and software computer infrastructure, that enabled the training of larger and more effective models.

An important concept in DL is *representation learning* [48]. Traditional ML techniques relied heavily on domain expertise for feature engineering and extraction from raw data, such as images. These engineered features are then used to train ML models. DL has dramatically shifted this approach: the model automatically discovers and extracts the necessary features for a given task, a process known as representation learning. The learned representations are hierarchical, meaning that the more abstract features are formed by composing simpler, multi-level representations [33]. Figure 1 shows how a deep learning model can capture the concept of a person in an image by assembling basic elements like edges, corners, and contours.

The basic element of Neural Network (NN)s is the artificial neuron, a biologically inspired model introduced for the first time in 1943 by McCulloch and Pitts [49]. A neuron is essentially a function that takes as input a number of inputs, combines them with a weighted sum, and then produces an output applying a non-linear function (named *activation* function). Neural networks are obtained by combining many neurons into a layered structure. The st deep learning architecture is the Multi-layer Perceptron (MLP), also known as fully-connected feedforward neural network, in which the neurons in each layer are connected to all the neurons in the next layer (Figure 2). From a mathematical point of view, these networks are essentially a chained composition of simple functions (the neurons).

Convolutional Neural Networks

A popular type of neural network is the Convolutional Neural Network (CNN), specifically designed to analyze data in the form of multiple arrays, such as 1D signals, 2D images, and 3D volumes or videos. CNNs are named after the basic operation they perform,

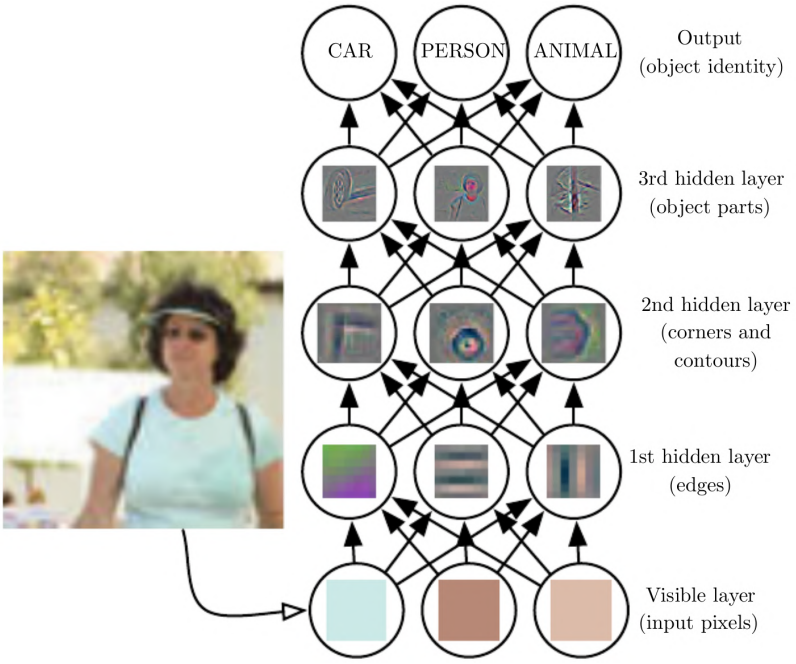


Figure 1: Illustration of hierarchical representation learning for a deep learning model. The crucial point is that these features are not manually engineered, but instead, they are automatically learned from data [33].

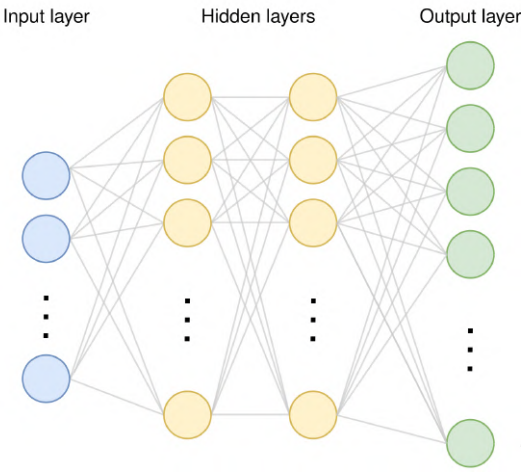


Figure 2: Example of MLP with two hidden layers.

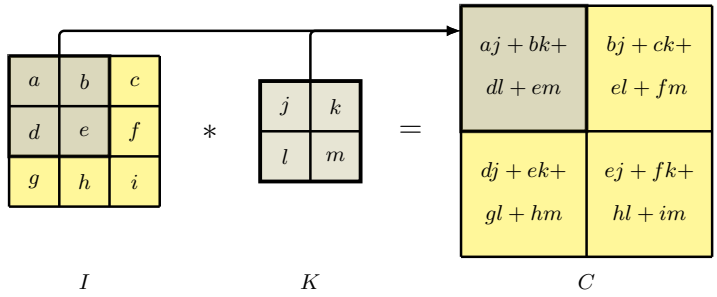


Figure 3: Example of convolution between a 3×3 image and a 2×2 kernel, that gives as output a 2×2 feature map [50].

the convolution[1]. This operation enables to extract features from data through specific filters (or kernels), that are learned during the training phase. For two-dimensional images, the convolution between an image $I(j,k)$ and a filter $K(l,m)$ can be expressed as [50]:

$$C(j,k) = \sum_l \sum_m I(j+l, k+m) K(l,m) \tag{1}$$

where $C(j,k)$ is usually named *feature map*. A schematic representation of the convolution operation for images is reported in Figure 3. Notice that in the case of 1D signals or 3D volumes an analogous definition can be made.

CNNs are typically made of three different types of layers [51]:

- *Convolutional Layers*. In these layers, various kernels are used to generate new feature maps, i.e. to extract specific patterns from an image.
- *Pooling Layers*. Pooling layers reduce the spatial dimension of the input image for the next convolution by aggregating nearby inputs with a summary statistic. This operation, also called subsampling or downsampling, is beneficial since it helps control overfitting, leads to faster convergence, and extracts space-invariant features [52]. The most common pooling strategies are average pooling and max pooling.
- *Fully Connected Layers*. After several convolutional and pooling layers, the abstract features extracted from the input data are given as input to fully connected layers that are used to make the final prediction.

An example of a typical CNN architecture is visible in Figure 4. The structure of CNNs incorporates several important concepts that align with our intuitive expectations. First, we want the network to detect

1 To be precise, most neural network libraries do not implement the convolution but a related function named cross-correlation, calling it convolution. For more detail see [33].

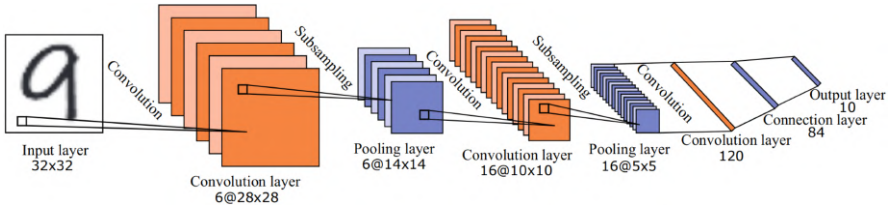


Figure 4: LeNet architecture, one of the first examples of CNN [34, 53].

patterns regardless of their location in the image or signal. This is achieved through *parameter sharing*, where each kernel operates independently of location and is applied uniformly across the entire image. Additionally, convolution is characterized by *sparse interactions*, meaning that each neuron in a layer is connected to only a small, localized region of the previous layer. This approach enables efficient processing and learning from data by focusing on local features rather than the entire input, thereby reducing computational complexity.

Training a neural network

To train a neural network, it is necessary to define a loss function that measures the error between the network's predictions and the desired output. The specific form of the loss function depends on the task. The most common ones are Cross Entropy Loss for classification and Mean Squared Error for regression. Once the loss function is defined, the training process involves minimizing the loss function. This is achieved through *gradient descent*, an optimization algorithm for finding a local minimum of differentiable multivariate functions that consists of iteratively adjusting the network's weights to follow the direction of the negative gradient. In practical applications, gradient descent in its vanilla version is never used. Instead, more advanced algorithms have been proposed to fasten and improve convergence, like Stochastic Gradient Descent (SGD) and Adam [54]. Gradient descent optimization relies on the gradient of the loss function with respect to the network's weights, which is computed through an algorithm called *backpropagation* [55].

Deep Neural Networks (DNNs) typically have a huge number of parameters and are prone to overfitting. To mitigate this, various regularization techniques have been developed. Among these, the most used are:

- *Dropout*. A certain percentage of neurons is deactivated during each training iteration [56]. This increases the network's robustness by preventing the network from becoming overly reliant on specific neurons.

- *Early stopping*. The data is divided into three parts: training, validation, and test sets. The training set is used to update the network's weights, while the validation set is used to test the

network in the training phase. The training is stopped when the validation error stops decreasing [57].

- *Weight decay* The squared value of the weights is added to the loss function, penalizing large weights and encouraging the model to distribute the importance more evenly. It is also known as L2 regularization.

1.2 RESEARCH QUESTIONS

This section introduces the diseases and the related research questions that will be addressed in the next chapters of the work.

1.2.1 *Atrial fibrillation*

AF is the most common cardiac arrhythmia, with an estimated worldwide prevalence of 59 million individuals in 2019 [2]. The incidence and prevalence of AF are increasing globally, to the point that it was defined as a 21st-century cardiovascular disease epidemic [58]. Projection studies indicate that the prevalence of AF will increase to 15.9 million in America by 2050 [59] and to 17.9 million in Europe by 2060 [60]. This trend is due to the aging population since age is one of the most important risk factors for AF. Other known risk factors for AF are obesity, smoking, hypertension, and diabetes mellitus, as well as cardiac pathologies such as myocardial infarction, Heart Failure (HF), rheumatic heart disease, and valvular disorders [61]. The global burden of AF is certainly underestimated, as AF is asymptomatic in approximately one-third of cases [62].

AF is associated with many complications, with stroke being the most relevant. Studies indicate that AF is associated with a 4- to 5-fold increased risk of stroke [63], and that among patients admitted to hospital with acute stroke in Europe and North America, 18 to 26% have pre-existing AF [64–66]. Other complications related to AF are HF, myocardial infarction, dementia, and chronic kidney disease [61]. Timely detection of AF and implementation of appropriate treatment, including rate and rhythm control medication and anticoagulant therapy, could reduce the frequency of AF-associated complications.

AF is a major public health burden, also associated with significant healthcare costs [67]. For effective AF management, accurate and timely detection of AF is essential. For this reason, a lot of recent research aimed to develop AF detection and prediction tools, including methods based on AI. For a comprehensive review of recent advancements in AI techniques for AF detection, prediction, and risk stratification, see the work by Salvi et al. [68].

The diagnosis of AF is most commonly done with the ECG, a non-invasive exam that records the heart's electrical activity (section 1.3.1).

AF is characterized on an ECG by the absence of consistent P-waves and irregular RR intervals [69]. From a clinical point of view, AF presentation can be classified on the basis of the temporal pattern of the arrhythmia. Recurrent atrial fibrillation occurs when a patient develops two or more episodes of the disorder, which could be paroxysmal or persistent: paroxysmal, if they terminate spontaneously within seven days, persistent if the arrhythmia continues requiring electrical or pharmacological cardioversion for termination. AF that cannot be successfully terminated by cardioversion, and longstanding (> 1 year), is named permanent [70].

The identification of patients at high risk of developing AF through routine, low-cost exams such as the ECG is of significant interest. The work presented in Chapter 2 addresses this by exploring different aspects regarding the development of ML-based clinical prediction models for AF.

1.2.2 *Dilated Cardiomyopathy*

Cardiomyopathies are a heterogeneous group of disorders of the cardiac muscle, associated with mechanical and/or electrical dysfunction that results in inappropriate ventricular hypertrophy or dilatation [71]. They can be divided into 5 different types: dilated cardiomyopathy, hypertrophic cardiomyopathy, restrictive cardiomyopathy, arrhythmogenic cardiomyopathy, and Takotsubo cardiomyopathy [72]. Dilated Cardiomyopathy (DCM) is defined by left or biventricular dilatation and contractile dysfunction in the absence of abnormal loading conditions and severe coronary artery disease.

DCM is a leading cause of heart failure and the most common reason for heart transplantation globally [73]. Its prevalence is estimated at 1 in 250 to 400 patients among those with heart failure, and 1 in 2500 in the general population, equating to roughly 40 cases per 100 000 individuals [72]. The annual incidence is 7 cases per 100 000 individuals. There are racial disparities in the incidence, while sex-related differences are less pronounced. This condition accounts for about 60% of childhood cardiomyopathies, with the highest incidence occurring in infants younger than 12 months. DCM is mainly caused by genetic mutations, that account for up to 35% of the cases [74]. The genes involved are the ones, among others, that encode cytoskeletal, sarcomere, and nuclear envelope proteins. The aetiologies for DCM include also infections, inflammation, autoimmune diseases, endocrine disturbances, and exposure to toxins like alcohol, cocaine, and methamphetamines.

Regarding the clinical manifestation, patients with DCM typically present with signs of HF, such as dyspnea, congestive edema, orthopnea, fatigue, and chest pain. Other DCM manifestations can be arrhythmias and Sudden Cardiac Death (SCD). The diagnostic investigations

of DCM include echocardiography, ECG exams, CMR, laboratory tests, and genetic testing. Once DCM is established, the treatment is directed at the major clinical manifestations of HF and arrhythmias. The prognosis is poor for DCM patients with a Left Ventricular Ejection Fraction (LVEF) $< 35\%$, a right ventricular involvement, a poor New York Heart Association (NYHA) functional class, and a poor hemodynamic status at cardiac catheterization [75]. A study with historical survival data from tertiary referral centers in adult patients with DCM [75] indicated a 1-year mortality of 25–30% and a 5-year survival rate of 50%, with SCD occurring in up to 12% of cases and accounting for 25–30% of all deaths.

Despite multiple phenotypic and genetic findings that have been suggested as possible prognostic features of HF and major arrhythmic events, no definite prognostic risk score has yet been validated in this field. Recently, AI has been used to characterize distinct phenogroups within the DCM spectrum, using second and third-level exams, such as CMR data and endomyocardial biopsy. Similar studies carried out only with data that can be collected upon first medical contact, like the ECG, are lacking. It is in this context that the work presented in Section 3.2 stands.

1.2.3 *Anomalous Aortic Origin of the Coronary Arteries*

Coronary artery anomalies are a rare form of congenital heart disease that encompass a wide spectrum of variants, each with a prevalence of $< 1\%$ in the general population [76]. A specific subset is the Anomalous Aortic Origin of Coronary Arteries (AAOCA), which is the abnormal origin or course of one or more coronary arteries that arise from the aorta. AAOCA is the second leading cause of SCD in young US athletes [77], and it also accounted for up to one-third of the deaths documented in autopsy reports of young military recruits who died during intense physical activity [78]. AAOCA comes with many different anatomical variations, some of them considered at high risk as associated with an anticipated higher risk of SCD (especially the ones with an interarterial course between the great arteries and an intramural course). Patients with AAOCA may exhibit a range of symptoms, though half are asymptomatic at the time of presentation. Notable symptoms include chest pain, dyspnea, palpitations, and syncope [79]. For a correct diagnosis of AAOCA, advanced image techniques like Coronary Computed Tomography Angiography (CCTA) or CMR are necessary. Beyond the diagnosis itself, it is crucial to identify the anatomy of the anomaly for a correct risk assessment.

Advances in imaging technologies and the increased use of screening protocols have significantly raised the number of adult patients diagnosed with AAOCA in recent years [76]. Although current guidelines offer general principles for diagnosis and treatment, the wide

range of anomalies and symptoms in AAOCA patients limits the applicability of these recommendations in clinical practice. AAOCA can be managed with a variety of interventions, and there is no consensus on the optimal treatment approach [76, 80]. According to the 2020 European Society of Cardiology (ESC) guidelines [81], surgical repair is the recommended treatment for AAOCA patients presenting with symptoms or evidence of stress-induced myocardial ischemia.

Timely and accurate detection of AAOCA is crucial for its optimal management. While AAOCA can be incidentally discovered during imaging studies conducted for other conditions, such as Coronary Artery Disease (CAD), it can still be missed if physicians are not actively looking for it. Therefore, there is an unmet need for automated tools to analyze CCTA images to reduce the risk of overlooking rare high-risk AAOCA. The study presented in Chapter 4 tries to fill this need, with the development of a fully automated AI-based screening tool for detecting and classifying AAOCA in 3D-CCTA images.

1.2.4 *Transthyretin Amyloid Cardiomyopathy*

Transthyretin Amyloid Cardiomyopathy (ATTR-CM) is an underdiagnosed progressive cardiac disorder caused by the misfolding of transthyretin amyloid protein, leading to the deposition of amyloid fibrils in the extracellular space of cardiac tissues [82]. The amyloid accumulation can result in abnormalities in atrioventricular conduction and stiffening of the myocardial tissue, which ultimately impairs cardiac function, leading to heart failure and impaired prognosis [82, 83]. In addition to the myocardium, amyloid fibrils can affect valve tissue, damaging endothelial cells and eventually causing calcification, particularly of the aortic valves, facilitating the development of Aortic Stenosis (AS) [82–84].

Recent studies have suggested that the coexistence of severe AS and ATTR-CM is more frequent than previously anticipated and associated with an increased risk of adverse events after Transcatheter Aortic Valve Implantation (TAVI) [85–88]. Although TAVI has been shown effective in this high-risk patient population, it is unlikely to achieve sustained improvement in symptoms and prognosis without addressing the underlying cardiomyopathy, for which reason timely diagnosis and subsequent treatment of ATTR-CM are central to the optimal patient management in this specific population [88]. Current ESC [89] and ACC/AHA [90] guidelines propose different algorithms for ATTR-CM diagnosis. Initial assessments typically involve clinical examinations, ECG, echocardiography, and CMR imaging to include or exclude potential patients based on specific symptoms [82]. For example, bilateral carpal tunnel syndrome and peripheral neuropathy as clinical features, low QRS voltage and pseudo infarct patterns in ECG, apical sparing, or increased atrial/RV wall thickness in echocardiog-

raphy could serve as red flags for ATTR-CM [82, 91–94]. While these modalities are useful in the preliminary evaluation, they are not specific to ATTR-CM; thus, a final diagnosis often cannot be based entirely on these results [82, 95, 96]. A definitive diagnosis of ATTR-CM may hinge on the histopathological confirmation or proof of a TTR mutation while always requiring confirmation of cardiac involvement, e.g., by significant cardiac uptake in scintigraphy [82, 97, 98]. Pathology and genetic testing are invasive and costly, while scintigraphy adds a significant financial and procedural burden, especially for severe AS patients undergoing TAVI who have already undergone extensive examinations [82–84]. Therefore, developing a non-invasive, financially viable method based on available data from preprocedural and routine data would be highly beneficial for detecting ATTR-CM. This challenge is addressed in Chapter 5.

1.3 DATA MODALITIES

In this section, we introduce the main data modalities that will be used in the following chapters to build the prediction and detection models.

1.3.1 *Electrocardiogram*

An ECG is a recording of the heart's electrical activity [99]. It is obtained through electrodes placed on the skin that detect the electrical signals resulting from the cardiac muscle's depolarization and subsequent repolarization during each heartbeat. The heart's electrical activity can be measured in different directions, termed *leads*. In a typical ECG, 12 leads are measured through 10 electrodes.

The central element in an ECG is the P-QRS-T complex, which represents the electrical signal of contraction and relaxation of the atria and ventricles. The typical shape of such complex is reported in figure 5, and its main components are [100]:

- P wave: atrial depolarization wave
- PR interval: distance from the beginning of the P wave to the beginning of the QRS complex
- QRS complex: it is made of three waves (Q, R, and S) that correspond to the same event, the ventricular depolarization
- QT interval: it represents the sum of depolarization (QRS complex) and repolarization (ST segment and T wave). It is necessary to correct it for heart rate (QTc)
- ST segment and T wave: the T wave, together with the preceding ST segment, is formed during ventricular repolarization

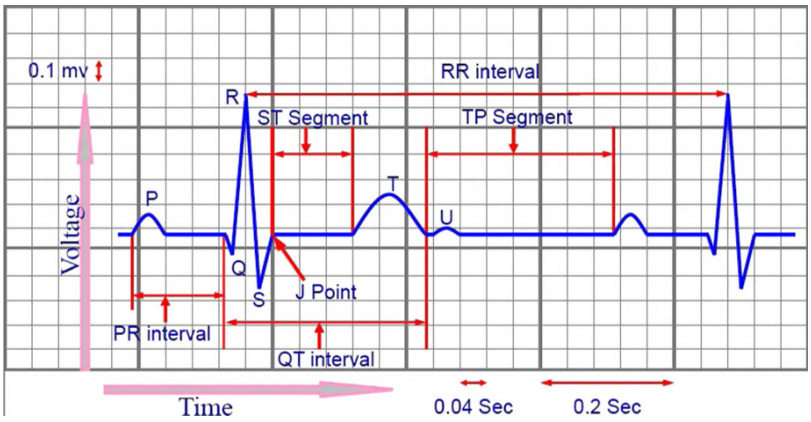


Figure 5: Standard fiducial points in the ECG (P, Q, R, S, T, and U) together with clinical features [101].

- RR interval: it's the *heartbeat* interval, and it corresponds to the time between the R peak of a heartbeat and the following/preceding heartbeat

The ECG is a widely used, simple, inexpensive, and safe diagnostic tool that has been a cornerstone since its introduction in medicine, more than 100 years ago [102]. It remains the most important technique for recording the electrical activity of the heart and diagnosing a wide range of heart diseases. The ECG is invaluable for diagnosing and evaluating active and passive arrhythmias, pre-excitation syndromes, channelopathies, interatrial and ventricular blocks, and acute ischemic events [103]. Additionally, the ECG is employed in various biomedical applications, including heart rate measurement, heartbeat rhythm analysis, emotion recognition, and biometric identification. Its non-invasive nature, combined with its diagnostic power, makes the ECG an indispensable tool in cardiovascular health. However, the ECG signal is complex to analyze, making its interpretation time-consuming even for experts. Therefore, computer-aided methods are necessary to alleviate the human burden and reduce errors caused by fatigue and inter- and intra-observer variability. In recent years, ML techniques have been applied to ECG data for the identification and classification of cardiovascular diseases, showing outstanding performance for certain tasks [104–106].

1.3.2 *Coronary Computed Tomography Angiography*

CT is an imaging technique that enables to obtain high-resolution images of the body by measuring different tissues' X-ray attenuations. It makes use of rotating X-ray tubes and detectors, and specific reconstruction algorithms to create cross-sectional images (also known as *slices*) of the internal structure of the body. Computed Tomography Angiography (CTA) refers to the imaging technique used to visualize

arteries and veins of the body. It is a contrast-enhanced CT, meaning that patients are injected with a radiocontrast substance that makes vessels easier to visualize. CTA produces detailed images that help detect blockages, aneurysms, dissections, and stenosis. This technique can be applied to various regions, including the heart. Specifically, CCTA is a specialized form of CTA that focuses on the coronary arteries of the heart [107]. CCTA is particularly challenging since the heart is a moving organ. For this reason, high temporal resolution and ECG-synchronized acquisition protocol are required. This means that during the acquisition the the patient is monitored by an ECG, and images are obtained at specific time points in the cardiac cycle, eliminating cardiac motion artifacts (Figure 6). This allows physicians to assess blockages in the coronary arteries, typically to diagnose coronary artery disease.

Over the past two decades, CCTA has become a critical diagnostic tool, influencing clinical guidelines and practices [108]. CCTA is mainly used for cardiac vascular disease detection, in particular, to exclude significant coronary artery obstruction in patients with low/intermediate probability of disease [107]. Other applications include triaging low-risk patients with acute chest pain (enabling safe and early discharge from the emergency department), guiding patients referral to catheterization laboratories, monitoring patients after coronary revascularization, and aiding in cardiovascular risk stratification [107, 108].

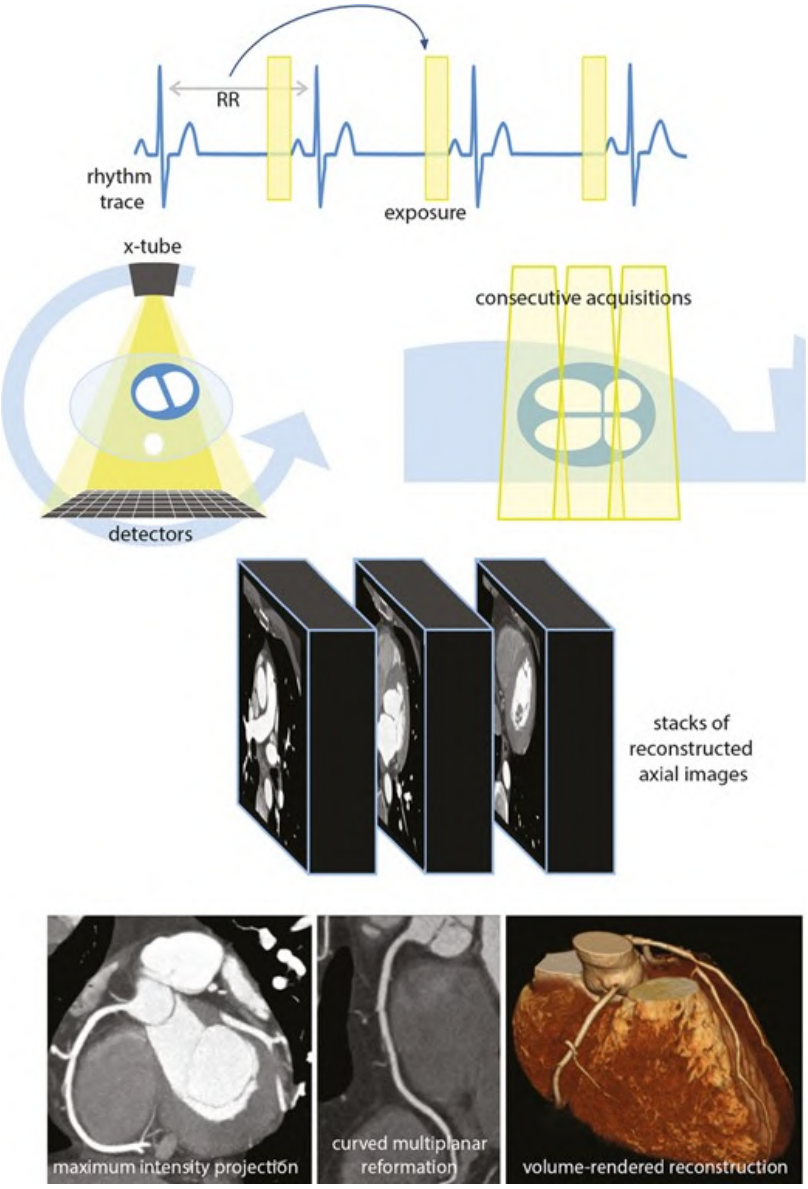


Figure 6: Using ECG-synchronization, the heart is scanned during the same pre-specified phase of contraction over several heart cycles. The consecutive stacks of axial slices are compiled into a single 3D image. [107].

2

ATRIAL FIBRILLATION PREDICTION

This chapter focuses on the development of risk prediction models for the prediction of AF using ECG data.

The first section investigates the discrimination and calibration performance of two machine learning models, with particular attention to the impact of class imbalance corrections during training. Atrial fibrillation prediction from ECG signals serves as the case study for this analysis. The second section addresses the underexplored topic of sample size determination for deep learning models. Using a learning-curve approach, it examines both the prediction and detection of atrial fibrillation from ECG signals. Finally, the third section extends the work on the prediction of AF from ECG data. The key advancements include developing the model within a survival analysis framework that accounts for competing events and integrating ECG signals with Electronic Health Records (EHR) data.

The research presented in this chapter is based on the following studies:

- **Giovanni Baj** et al. "Comparison of discrimination and calibration performance of ECG-based machine learning models for prediction of new-onset atrial fibrillation." In: *BMC Medical Research Methodology* 23.1 (July 2023), p. 169. DOI: 10.1186/s12874023-01989-3
- Emanuele Giusti, Annamaria Porreca, **Giovanni Baj**, Ilaria Gandin, Giulia Barbati, Daniela Pacella, "Sample size determination via learning curves for AI models: an application to deep learning algorithms for diagnostic and prediction tasks". Submitted to *Artificial Intelligence in Medicine*.
- **Giovanni Baj** et al. "Deep Learning Survival Model to Predict Atrial Fibrillation From ECGs and EHR Data." In: *Progress in Artificial Intelligence*. Springer Nature Switzerland, 2023, pp. 222–233. DOI: 10.1007/978-3-031-49011-8_18

2.1 EFFECTS OF CLASS IMBALANCE CORRECTIONS ON ML MODELS PERFORMANCE

2.1.1 *Introduction*

In the last few years, there has been a growing interest in the potential diagnostic value provided by ECG signals. ECG waveform is one of the most extensively studied physiological signals to evaluate the condition of the heart, in which several waves as P, R, and T, are key to determining the type of rhythm. The interpretation of ECGs is complex and requires inspection by highly trained clinicians. However, numerous studies have shown that computer-aided methods based on ECG data represent a promising tool for the analysis and identification of cardiovascular diseases [104].

One example is the prediction of AF, the most common supraventricular arrhythmia in the general population (Section 1.2.1). AF is a relevant risk factor for stroke, however, it is often asymptomatic and not recognized. Thus, the identification of patients at high risk of future development of AF represents a major challenge. The way AF detection and prediction are evolving with the availability of new predictive tools is well described in a review carried out by Siontis et al. [109]. The development of tools to predict AF from routine and low-cost exams such as ECG would be an important step toward the active targeting of patients at risk, a task for which clinical risk scores and electronic health record-based tools have shown limited power [110].

The 12-lead ECG is a rapid, cost-effective cardiological exam that is routinely performed at different levels of point-of-care, from hospitals to clinics and ambulatory centers, generating a massive number of digital traces. As for other types of Big Data in the healthcare context, a major role in their analysis may be played by AI systems, which can be easily fed with hundreds of thousands of observations. Two main approaches can be distinguished for the development of diagnostic models based on ECG. One approach involves the analysis of ECG features. Automated ECG interpretation is not a new concept, and algorithms that provide ECG interpretations have been around for a long time (in many cases code is proprietary and not disclosed). Such computer programs usually work in separate stages that include signal pre-processing, beat identification, correction, computation of average beats, and identification of fiducial points from which ECG measurements are extracted. Such measurements rely on knowledge-driven markers (like QRS, ST-segment elevation, T-wave changes) reflecting the clinical knowledge of heart activity, and can be then used to define criteria and rules for a diagnostic evaluation by physicians. In addition to human evaluation, in the last years ECG features, which can vary in number and type depending on the program employed,

have been used to feed ML methods for tabular data to derive a diagnostic model [111–114].

The second approach consists of developing end-to-end prediction models that do not require feature extraction. This strategy involves the direct analysis of the digital ECG waveform to obtain the probability of a specific class in the classification of interest and DL based neural networks have demonstrated to be able to achieve good results. Despite DL models being black boxes and requiring the application of explainability techniques to investigate their prediction mechanism, this AI method for ECG analysis is being increasingly explored for its ability to detect subtle and non-linear interrelated variations along the signal [105]. The most common DL architectures used for analyzing ECGs are Convolutional Neural Network (CNN)s, a specialized kind of neural network for pattern recognition in time series and image data (Section 1.1.2). These networks can be thought of as having two sequential components: in the first layers, a set of convolutional filters allows us to extract patterns and key features from the signal, while in the second part, these extracted features are combined and used to make a prediction. Notice that the specific weights of the filters to be applied and the relative features extracted are automatically learned by the network in the training process. It has been recently shown that the performance of a CNN in classifying arrhythmia from ECG can exceed that of cardiologists with average experience [8, 115]. Besides this classification task, CNNs have already shown good performances in predicting the new onset of AF (see Raghunath et al. [12] for AF prediction within 1 year, and Attia et al. [11] for the identification of electrocardiographic signature of AF immediately prior to diagnosis). All these works reported quite good values of discrimination accuracy, but no information was available about the calibration of the estimated probabilities.

Note that the diagnostic/classification task is quite different from the prediction task in epidemiological studies: classification is best used to identify the presence of an outcome/condition in the context of case-control studies. On the contrary, in the context of cohort studies when subjects are selected as initially free from the outcome and are then followed in time until they will (or will not) develop the outcome, usually observed in a minority of subjects, modeling tendencies (i.e., probabilities) is key [116]. The common approach of balancing events/non-events cases before applying ML/DL algorithms, based on the perception that this procedure can improve the performance, seems not advisable in the prediction context [117]. The consequence of balancing could be that the algorithm trained to "predict" a 1/2 incidence of events will not be applicable to a population with a 1/1000 incidence. Subsequent calibration procedures are then needed in order to correct this issue [117]. Since the low incidence of new-onset

AF in our population, the possible impact of balancing was an issue that we wanted to explore in the context of AF prediction.

The main goal of the present research was the development of a predictive model for a binary outcome based only on ECG information by comparing different methods: an ML algorithm on signal features and a DL approach on raw signals. Penalized logistic regression was used as a benchmark method. In this framework, AF represents a case study and this research does not claim to propose a prediction tool suitable for clinical practice. Instead, our effort is aimed to extensively analyze the performance of the two approaches in terms of discrimination and calibration taking into account varying sample sizes and degrees of balance between the events and the censored cases. In particular, our research was based on different hypotheses: a) DL models based on the raw ECG signals could potentially outperform algorithms working on ECG features when the training set is large; b) the use of under-sampling to handle class imbalance does not improve discriminative performance and could instead produce miscalibrated predictions.

2.1.2 *Background*

Calibration

As it is well known, discrimination refers to a model's ability to rank patients by risk, typically measured with the Area Under the Receiver Operating Characteristic Curve (AUC). Another critical but often overlooked metric in risk prediction models is calibration [116]. Calibration assesses the agreement between estimated risk and observed incidence, a fundamental aspect when applying models in clinical settings. Indeed, decisions are frequently based on risk, so the model's probability estimates should be reliable for optimal decision-making. Miscalibration can reduce a model's clinical utility and, in some cases, even make it harmful [118]. Poor calibration can arise from differences in patient characteristics and disease prevalence between the development cohort and the application cohort, or from statistical overfitting, which occurs when a model is too complex for the available data.

Calibration can be assessed visually using a *reliability diagram*, where the estimated risk is plotted against the observed proportion of events. For a well-calibrated model, the calibration curve should align closely with the diagonal line. Calibration can also be evaluated quantitatively using various methods [117], including the Cox's intercept and slope [119], the Brier score [120], and the Hosmer-Lemeshow test [121]. In the works reported in this thesis, we chose to evaluate calibration with the Integrated Calibration Index (ICI), a method recently proposed by Austin et al. [122]. Similarly to Cox's method [119], the

ICI is based on a graphical assessment of calibration, where the observed binary outcome is regressed on the predicted probability of the outcome using a locally weighted least squares regression smoother (i.e., the Loess algorithm). Then, ICI is given by a numerical summary of calibration, computed as the weighted average of the absolute difference between the smoothed regression line and the diagonal line of perfect calibration, where the weights are given by the density function of the predicted probabilities. For a perfectly calibrated model, ICI takes value 0, and in general the higher the ICI, the less the model is calibrated. We opted not to use the more employed Cox' slope and intercept because these metrics could be equal to their ideal values of 0 and 1, respectively, while deviations of the calibration curve from the line of identity can still occur [122]. Moreover, the ICI can be easily generalized to evaluate prediction models in survival settings, even in the presence of competing risks, which is the focus of the work in Section 2.3.

Imbalance corrections

When developing clinical prediction models for binary outcomes, it's common to encounter situations where the event of interest occurs in significantly fewer than 50% of cases, resulting in what is known as *class imbalance*. In the ML community, class imbalance has been recognized as a challenge, and various solutions have been proposed to address it [123–125]. One straightforward approach is to artificially balance the dataset during model training. This can be achieved by oversampling the minority class, either by repeating existing positive samples or by creating new synthetic samples, or by randomly undersampling the majority class. However, recent research has shown that applying imbalance corrections in the development of classical statistical models can lead to strong miscalibration [126]. In this section, we conduct a similar study within the context of ML models.

2.1.3 *Methods*

Data

The dataset used for AF prediction originates from the Trieste Observatory of Cardiovascular Diseases [127], established in 2009 to integrate administrative and clinical data sources for epidemiological studies based on real-world populations [128]. The Observatory combines data from administrative and clinical sources. The administrative data comes from the Friuli Venezia Giulia (FVG) Regional Epidemiological Repository (RER), an EHR system that includes registries of births and deaths, hospital discharge records, laboratory tests performed in public hospitals, and public drug distribution records. It covers all beneficiaries of the Italian national healthcare system resid-

ing in the FVG region, approximately 1.2 million people. The clinical data is sourced from C@RDIONET, a cardiological electronic chart that includes medical information collected by cardiologists during routine clinical practice in the FVG region. The integrated database specifically covers the population from the municipalities of Trieste and Gorizia, comprising approximately 240.000 inhabitants.

Regarding the ECG dataset, it includes all ECG exams acquired at the Cardiovascular Department of Azienda Sanitaria Universitaria Giuliano Isontina (ASUGI) in Trieste from February 2007 on. ECGs were recorded at a frequency of 1 kHz using the Mortara™ devices ELI230 and ELI250 and then resampled at 500 Hz for computational reasons. By linking the ECG exams with the FVG RER, we could integrate them with C@RDIONET and then identify a cohort of patients without AF history for the prediction of the new onset of AF. The AF event was defined by linking information from 4 different sources: reports from emergency access or cardiological visits, discharge codes in case of hospitalizations, and ECG reports. For each patient, the first AF diagnosis (or atrial flutter) found in one of these data sources was taken as the first AF event. We excluded all patients with an AF event before 2007 or with paced rhythms (i.e., implanted with a pacemaker, with an implantable cardiac defibrillator, or treated with cardiac resynchronization therapy). Subjects with an AF diagnosis at the first ECG exam or with the AF-event date missing were not included in the analysis.

For this study, we included all subjects aged $>$ 30 years with at least one standard 10-second, 12-lead ECG acquired at the cardiovascular department of ASUGI, between February 2, 2007, and December 31, 2020. The data exclusion flow chart is reported in Figure 7. For patients without any AF event in the observation period, we extracted all available ECG exams, while for patients that developed AF, we used all ECGs recorded before the first AF event within a temporal window of 5 years. Note that censored cases, i.e., subjects that did not develop AF, had a minimum follow-up of 5 years required by design. Each ECG was associated with a set of morphological features, automatically extracted by the Mortara devices at the ECG recording. We had access to these features through the cardiological e-chart C@RDIONET. The unit of observation was the ECG signal. Each ECG was labeled 1 if the corresponding patient will develop AF within 5 years, and 0 otherwise.

Models' development

The two approaches under study were a deep Convolutional Neural Network (CNN) and an eXtreme Gradient Boosting (XGB) model. A penalized Logistic Regression (LR) model was used as benchmark. For all models, the task considered was to predict the probability that a patient will develop AF within five years.

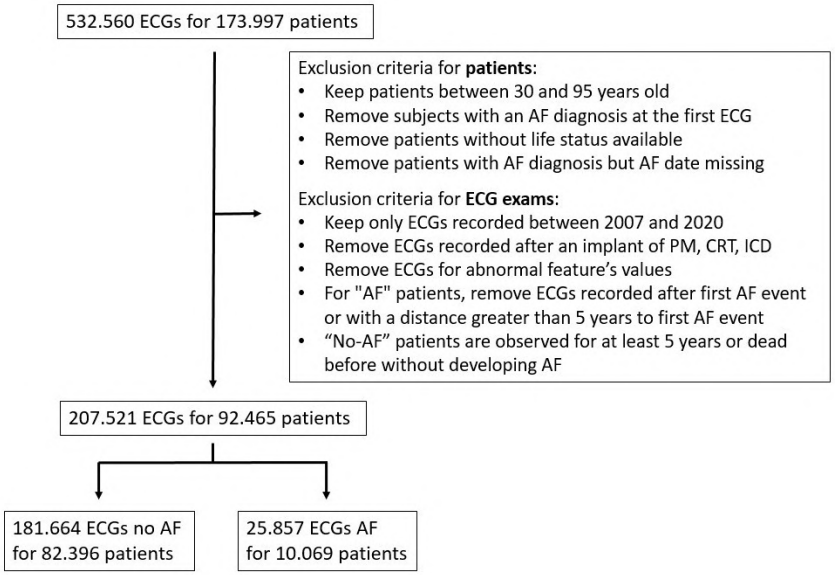


Figure 7: Flow chart of the study cohort.

The CNN takes as input the resampled ECG signal, which is a 12×5000 matrix (i.e., 12 leads by 10-second duration sampled at 500 Hz). The architecture of the CNN is the one used by Scagnetto et al. [129] for AF prediction, which was originally proposed by Goodfellow et al. [130] for a similar purpose, i.e., to classify single lead ECG waveforms as either Normal Sinus Rhythm, AF, or Other Rhythm. The network is composed of 13 blocks, each of which comprises a 1D convolution along the time domain, batch normalization, ReLU activation function, and dropout. Notice that, in the computation of convolutions, all channels are used simultaneously, thus cross-lead correlations are automatically leveraged by the model. In blocks 1,6 and 11 there is also a max-pooling layer between Rectified Linear Unit (ReLU) activation and dropout. After the convolutional blocks, there are a global average pooling layer and a soft-max layer, in order to obtain normalized probabilities. A sketch of the architecture and the most relevant hyperparameters are reported in Figure 8. To train the model we used the cross-entropy loss function and AdamW optimizer [131], with a learning rate of 10^{-3}.

The XGB and LR models take as input the wave morphology's features extracted from the ECG signal by the Mortara devices. These features include the onset and offset of P and T waves and the QRS complex, the PR and corrected QT intervals, P, T, QRS axis, and the cardiac frequency.

To tune the XGB's parameters, we performed a randomized search over parameters, as described hereafter. For each hyperparameter that we decided to tune, we specified a uniform distribution over the possible parameter values range. Then, we generated a candidate set-

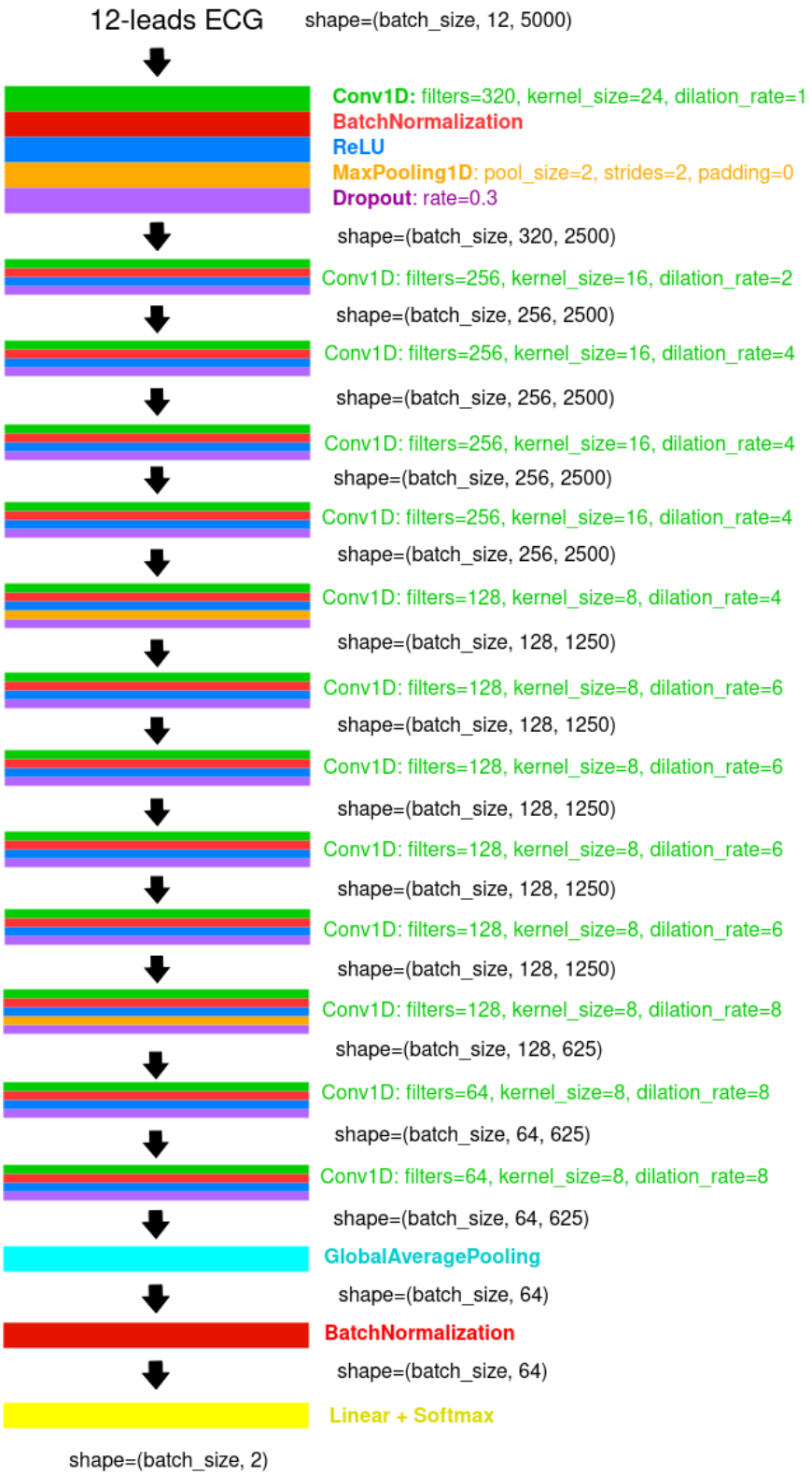


Figure 8: CNN architecture.

ting of parameters by sampling the pre-specified distributions and we evaluated its performance with a 5-fold cross-validation. These steps are repeated 10^5 times. Finally, the best-performing parameters setting (in terms of AUC) was kept. In this process, we included a set of ECGs (approximately 50 000) solely used for hyper-parameter tuning and not in training/test phases. More details about hyperparameter tuning are reported in the appendix A.

In the LR model, we applied an L2 regularization term to reduce overfitting. Therefore, the only parameter of the model is the strength of the regularization term, which we tuned with the same procedure followed for XGB hyperparameters tuning.

The CNN was implemented with `PyTorch` framework [132] version `1.12.0`, while for XGB and LR models we used `Scikit-learn` `1.0.2` implementations [133]. Python's version used was `3.10.5`. All the code used for this study can be found in the GitHub repository `https://github.com/giovabaj/ecg-cnn-xgb-lr`.

Models' evaluation

To assess the ability of the models to discriminate between patients developing/not developing AF, we used the AUC, which is a robust metric of model performance for binary classification, even in the case of imbalanced datasets. Higher AUC values correspond to better performances, with perfect discrimination represented by an AUC value of 1 and an AUC of 0.5 equivalent to a random guess. To evaluate the models' calibration, we computed the ICI (Section 2.1.2).

To evaluate the variability of the performance of the trained models, we performed a 10-fold cross-validation. In the case of the CNN, 8 sets were used to train the model, 1 to evaluate the model during training and apply early stopping, and the last one to test the model performances on unseen data. Regarding the XGB and LR models, data was split into training and test with a 9:1 ratio. In the process of splitting data into folds, in the case of patients with multiple ECGs we ensured that each patient was present only in one between training, validation, and test sets. This is because intra-patient ECGs show a higher degree of correlation with respect to inter-patient ECGs. Thus, without taking into account this detail, models' performances would be overestimated. We also made sure that the fraction of positive samples in each fold was as similar as possible to the overall fraction.

Experimental setting for varying sample sizes/balance

To investigate the dependence of models' performances on the sample size, we trained the three considered models with increasingly bigger subsets of the dataset. The sizes considered are 1000, 2000, 5000, 10 000, 20 000, 50 000, 100 000 and 150 000 ECGs. The remaining 57 521 ECGs were used to tune hyperparameters for all models.

Another aspect we investigated was imbalance corrections, again for increasingly bigger sample sizes. We repeated the training process described above, but this time balancing the two classes in the training set by Random Under Sampling (RUS), which consists of eliminating a random set of negative ECGs in order to equalize the number of ECGs in each class [124].

To study the effect of class imbalance corrections on models' performances, we considered a fixed sample size (100 000 ECGs), and we trained the three models with different balancing levels of the training set. The levels considered are 12.5% (corresponding to the original positive fraction), 25%, 37.5%, and 50% (perfectly balanced training set). We stress that the test sets used to evaluate the models have always the original positive fraction (12.5%). As before, the method used to balance the training set was RUS, and to estimate models' variability we performed a 10-fold cross-validation with the same approach described above. Notice that we decided to use a sample size of 100.000 ECGs since we observed that none of the three models showed a substantial improvement with a training size larger than this.

2.1.4 *Results*

The final dataset includes 207 521 ECGs, associated with 92 465 subjects. The number of events (i.e. new onset of AF) is 25 857, corresponding to 12.5% of cases. See Table 1 for a descriptive snapshot of the population. Note that the statistical unit for the prediction algorithms is the ECG signal. Compared with censored subjects, patients developing AF were older and more frequently male. These results are not surprising since increasing age is a prominent AF risk factor and the prevalence of AF is lower in women vs. men in most of the real-life study cohorts [134, 135]. Note that we did not include demographic characteristics in the analysis since the objective was to investigate the specific ECG contribution to the prediction. No other remarkable clinical differences are observed in the ECG features.

Models' evaluation results

In Table 2 we report AUC and ICI values and corresponding 95% Confidence Interval (CI) for the three models trained with the biggest sample size considered (150 000 ECGs) and with the original event ratio (no imbalance corrections). We can see that from a discrimination point of view, the CNN model is the best-performing model, with an AUC of 0.799. XGB model is the one with intermediate performance (AUC of 0.74), while LR shows the worst performance (AUC of 0.68). As for the calibration, it can be noticed that there are no substantial differences in the performance of the three models; XGB is the best-performing model with an ICI of 0.008, while the other two models

Table 1: Descriptive features of the dataset. For all the numerical variables median and (1st, 3rd quartile) are reported. We compared "Censored" and "Event" populations with Mann-Whitney and Chi-squared tests, respectively for continuous variables and gender. All comparisons were significant (p-value < 0.001).

	Censored	**Event**	**Overall**
Age (years)	65 (52, 75)	74 (67, 80)	67 (54, 76)
Gender (Male, %)	49	58	50
P axis (degrees)	58 (43, 69)	60 (42, 73)	58 (43, 69)
P onset (msec)	290 (269, 307)	274 (246, 295)	288 (266, 306)
P offset (msec)	407 (388, 422)	391 (361, 413)	406 (385, 422)
PR interval (msec)	163 (148, 182)	176 (157, 199)	164 (149, 184)
QRS axis (degrees)	37 (1, 64)	16 (-22, 53)	35 (-2, 63)
QRS onset (msec)	453 (449, 458)	451 (445, 457)	453 (449, 458)
QRS offset (msec)	550 (543, 558)	551 (545, 563)	550 (543, 558)
QT interval corrected (msec)	408 (395, 424)	420 (404, 439)	409 (396, 426)
T axis (degrees)	54 (36, 68)	59 (34, 78)	55 (36, 69)
T offset (msec)	841 (820, 863)	853 (829, 878)	842 (821, 865)
Heart rate (beats/min)	71 (62, 80)	69 (61, 78)	70 (62, 80)

Table 2: Performances in discrimination and calibration of the three models.

	CNN	XGB	LR
AUC	0.799 (0.794, 0.805)	0.738 (0.732, 0.744)	0.683 (0.678, 0.688)
ICI	0.014 (0.01, 0.018)	0.008 (0.006, 0.010)	0.014 (0.013, 0.015)

show higher ICI values. In terms of 95% CI, the lower bound of CNN and LR correspond to the upper bound of XGB.

Results for varying sample sizes/balance

In Figure 9, we show the dependence of AUC on the sample size for the three proposed models, both in the imbalanced (Figure 9A) and perfectly balanced (Figure 9B) cases. We can notice that the model that is most affected by the sample size is the CNN: for small samples, the discriminative performances are very low (lower than 0.70), but above 10 000 samples the DL model significantly outperforms XGB and LR, reaching an AUC of 0.80 in the imbalanced case. On the other hand, XGB and LR's discrimination does not change significantly increasing the sample size, while the most visible effect is the greater variability for small sample sizes, as obviously expected. For these two models the maximum AUC values, obtained with the biggest sample size, are respectively 0.74 and 0.68. Another aspect to note is that balancing the training set with RUS to an event ratio of 0.5 (same number of AF cases and censored samples) does not improve discrimination in any of the models considered, also not for small sample sizes.

In Figure 9 it is also reported the ICI as a function of the sample size. Figure 9C represents the case where no imbalance corrections were introduced, and we can see that increasing the sample size has the effect of reducing the ICI (i.e., it improves calibration), for all the three models under study. In this setting, ICI values range from 0.06 for smaller sample sizes, to approximately 0.01 for the biggest sample size considered. When RUS is applied to balance the training set (Figure 9D), ICI takes higher values, indicating that models are worse calibrated. The effect is very strong for XGB and LR (ICI values between 0.30 and 0.35 for all the sample sizes considered) and slightly weaker for the CNN model (ICI values between 0.1 and 0.2), but still evident, especially if compared with the imbalanced case.

As regards the effects of imbalance corrections using different event ratios and a fixed size of 100 000 ECGs, we found that XGB's and LR's discrimination capabilities show very little dependence on the balancing level introduced. This is evident in Figure 10A, where AUCs are reported for the three models as a function of the event fraction in the training set. Indeed, it can be noticed that XGB and LR models

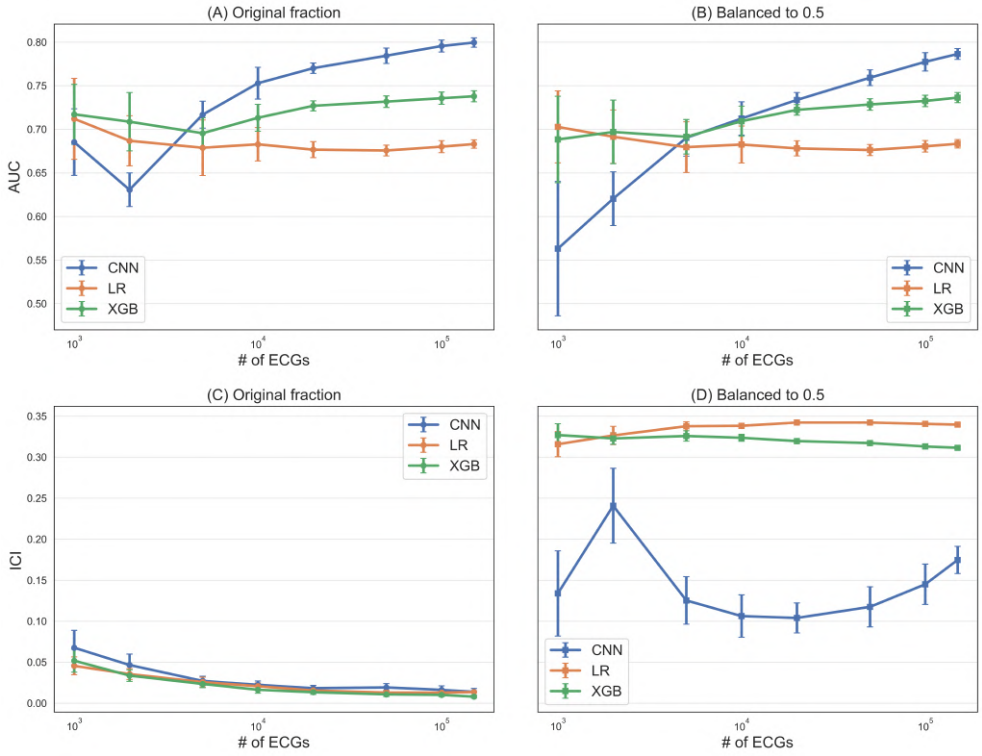


Figure 9: **A** AUC values for the varying sample sizes (original event fraction in the training set). Error bars represent the 95% CI around the mean. **B** AUC values for the varying sample sizes (perfectly balanced training set). **C** ICI values for the varying sample sizes (original event fraction in the training set). **D** ICI values for the varying sample sizes (perfectly balanced training set)

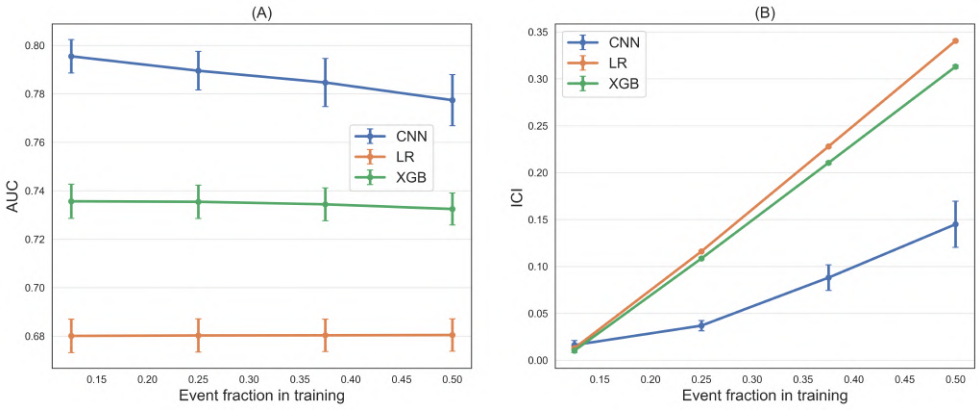


Figure 10: **A** AUC values for the varying event fraction, obtained by balancing the training set with RUS. Error bars represent the 95% CI around the mean. **B** ICI values for the varying event fraction, obtained by balancing the training set with RUS

show nearly constant AUC values, respectively of 0.74 and 0.68. As regards the CNN model, also in this case RUS does not allow to get better discriminative performances, rather AUC slightly decreases as we increase the level of imbalance corrections, approximately from 0.795 to 0.777. Moving to calibration (Figure 10B), the effect of balancing the training set was very clear: increasing the ratio of positive samples with RUS leads to higher values of the ICI, i.e. to less calibrated models. The effect is very strong for XGB and LR, where ICI values grow linearly from 0.01 to 0.3, and a little weaker for the deep learning model (ICI values from 0.01 to 0.15), but still evident.

2.1.5 *Discussion*

In this study, we investigated the use of ECG signals for the development of a predictive model for new-onset AF. This is a critical medical task because of the high prevalence of AF particularly in the elderly population and the importance of an early diagnosis of AF for prompt prescription of effective treatments to prevent stroke and systemic thromboembolism. Two approaches were considered: first, an ML model based on the set of ECG features extracted from the ECG and accessible to clinicians; second, the analysis of the digital ECG traces using deep learning techniques, in a setting of end-to-end analysis. In addition, a logistic regression model based on ECG features was estimated to provide a benchmark for the comparison of results. As for the analysis of ECG features, for large sample sizes, the XGB algorithm produced a model that outperformed the benchmark in terms of discrimination ability. In particular, the XGB and LR models appeared almost equivalent when the number of observations was lower than 10 000, but for larger sample sizes XGB demonstrated a clear increase in the level of discrimination, resulting however constant in further

enlargements of the dataset. In contrast, the CNN model showed a discriminative performance highly dependent on the sample size: to reach a satisfactory result, the DL model required at least 10^4 observations, but for every further increase of the size we observed a correspondent improvement in discrimination. In terms of calibration, no major differences were detected across models when the original fraction of cases was used. In general, we observed better-calibrated predictions for increasing sample sizes. Our results may suggest that the choice of the approach in the analysis of ECG should take into account the amount of data available for the training, preferring more standard models for small datasets, and indicate the well-known ability of DL methods to leverage massive datasets. The second part of our analysis was focused on the effect of undersampling on models' calibration. This aspect of the study was stimulated by a recently published work by van den Goorbergh et al. [126] where authors examined the effect of imbalance correction on the performance of standard and penalized (ridge) LR models in terms of discrimination, calibration, and classification. When developing prediction models for a binary outcome with high class imbalance, undersampling is a standard technique for mitigating the difference in class frequencies in the training phase, with the aim of improving the model's performance. We analyzed the results of models obtained with different levels of balancing ratios and failed to detect an improvement in discrimination, leading to even worse results in the case of CNN. Besides, imbalance correction caused miscalibrated predictions. Our results are in line with the findings of van den Goorbergh et al. and extend their note of caution in using methods for class imbalance correction in the case for XGB and CNN models. We observe that in our study the CNN resulted more robust compared to XGB and LR to the calibration worsening caused by the imbalance correction, a counter-intuitive finding with respect to what was observed by Gou et al. [136]. Concerning the relative performance of our CNN approach with respect to the recent literature that investigated new onset of AF, Attia et al. [11] considered a set of 649 931 12-lead ECGs of patients $\geq$ 18 years and applied CNN to identify the electrocardiographic signature of future AF developed within one month from ECG examination (8.4% of the cohort). They obtained a very accurate model (AUC 0.90 [0.90–0.91]), but the sample size and the time-frame prediction period was clearly very different from ours. Another relevant study was carried out by Ragunath et al. [12], in which authors analyzed 1.6 M 12-lead ECGs from patients aged 18 years or older in order to identify individuals at risk of developing AF within 1 year. Training a CNN using only ECG traces as input, they were able to predict the new onset of AF with AUC of 0.83 (95% CI, 0.83 - 0.84). Although the sample size and observational period are different from ours also in this case, the per-

formance is comparable with our findings (Table2). No measures of calibration were reported in those works.

Our study has some limitations. First, we could not validate our findings in an external validation cohort that represents one of the most critical steps in the development of machine learning models in medicine, a context where internal validation is not considered sufficiently conservative [137]. Second, for AF subjects we only considered ECG exams no further than 5 years before the date of AF diagnosis. We set such constraints because based on clinical knowledge, AF individuals are unlikely to show predictive signs of the condition earlier than 5 years. The methodological choice is also in line with previous clinical scores and predictive models that are usually evaluated at a time horizon of 5 years of follow-up [110]. Third, in order to simplify the prediction task, we did not take into account the time-to-event in disease onset. Very recent research carried out by Khurshid et al. [138] has highlighted the potential of CNN for the prediction of the time-to-incident AF and obtained accurate predictions (5-years AUC 0.823 [95% CI, 0.790 - 0.856]). One of the advantages of the time-to-event data is the possibility to evaluate the accuracy of the model for any time frame from the baseline. Another possible limitation was the choice of the method to correct the class imbalance, as RUS is a very naive approach. The main obstacle here was to deal with entire signals. For example, a commonly used method that has shown good results in various applications is the Synthetic Minority Oversampling Technique (SMOTE) [139]. SMOTE is an oversampling approach that creates new, synthetic samples interpolating the original minority class samples. This method and its variations were developed for tabular data, but an extension in the case of signals is not straightforward. Some methods to generate synthetic ECG signals were recently proposed [140–143], but it was out of the scope of this work. Finally, the fact that only standard ECG features were used for the XGB approach is a clear limitation, considering that several ECG engineered features were shown to be highly predictive for AF detection [111] and AF risk prediction [144, 145]. We expect that including this kind of features engineered from the ECG signal could improve XGB performances. However, we want to highlight that we limited on purpose to the features automatically extracted by electrocardiographs since we wanted to consider a setting as simple as possible, where only the ECG exam is required, so that the prediction process can be easily automated without the need for feature engineering by experts.

2.1.6 *Conclusions*

The deep learning model under study showed a discriminative performance highly dependent on the sample size, outperforming the two approaches considered based on the signal's extracted features

only above a certain sample size threshold. This result suggests that the choice of approach in the analysis of ECG should be based on the amount of data available, preferring more standard models for small datasets.

Imbalance corrections with a random undersampling approach did not lead to better discrimination performance, but rather to an evident drop in models' calibration. This finding indicates that imbalance correction methods should be avoided when developing clinical prediction models.

2.2 LEARNING CURVE FOR DL

2.2.1 *Introduction*

Several factors can explain the widespread adoption of AIand machine learning in healthcare: first of all, these models, especially when pre-trained via transfer learning or trained on large datasets, can achieve generally higher accuracy on diagnostic and prognostic tasks; also, they are able to integrate multimodal data efficiently; finally, they do not require strong statistical assumptions. During the study design phase, integrating AI models requires the addition of steps such as the specification of the AI model, the definition of the AI architecture, the choice of the evaluation measure, and, more importantly, Sample Size Determination (SSD) [146]. The recent publication of extended guidelines for the reporting of AI in clinical trials [147, 148] and in all types of studies [149, 150] has confirmed the necessity of adopting a rigorous approach to the determination of the sample size requirements. Neural Networks and Random Forests, among others, have been shown to be more data-hungry than traditional statistical models such as logistic regression [151] and several studies demonstrated the inadequacy of the "factor 10" sample size rule applied to either the number of predictors or network weights [152, 153]. A methodology for SSD that has been recently gaining attention is fitting the learning curve of the classifier with an inverse power law function [154]. The algorithm described by Figueroa et al. [155] allowed authors to predict the performance of several Support Vector Machines (SVM) classifiers using weighted Non-linear Least Squares (NLS) optimization for fitting the learning curve. A recent implementation by Dayimu et al. [156], instead, tried to predict the performance of Elastic Nete, SVMs, Random Forests, and Gradient Boosted Trees using both NLS and Gaussian Process (GP) optimization methods. Additionally, the study extends Figueroa's implementation to transfer learning algorithms. However, to date, these methods have never been applied to the fitting of a learning curve for DL algorithms.

2.2.2 *Objectives*

The present study aims to provide an application of both the weighted NLS and GP method for fitting the inverse-power law form of the learning curve of a CNN trained on a task of detection and prediction of AF using ECG data. Fitting the classifier's learning curve allows us to estimate the sample size required to reach a pre-specified performance and its expected plateau.

2.2.3 *Background*

SSD refers to the process of estimating the optimal number of participants needed for a study to achieve statistically significant results, balancing between accuracy and resource efficiency [157]. Although recent reporting guidelines for AI-based clinical models emphasize the importance of a rigorous approach to SSD, many ML still neglect this step [158]. In this context, the work presented in Section 2.2 demonstrates a practical application of SSD for deep learning models, using a *learning-curve* approach.

A learning curve represents the relationship between a model's performance and the number of samples in the training set [159]. Once this curve is estimated with a relatively small number of data points, it can be used to predict how many samples are required for the algorithm to achieve a specific performance threshold or a sufficiently low generalization error. This is particularly important in medical applications, where collecting and labeling data for model training can be both time-consuming and expensive.

To estimate the learning curve in practice, the model is trained on datasets of varying sizes, and its performance is measured as a function of the training set size. Then the relationship between training data size and performance is modeled using a pre-defined function, typically a power-law function [159]. Once the curve is fitted, it is possible to extrapolate the model's performance at larger sample sizes.

2.2.4 *Methods*

The dataset used for the detection task comes from the 2020 Physionet challenge [160], a multi-source dataset comprising 88 168 12-lead ECGs labeled for different types of abnormalities. Signals have different time lengths, but only the ones with a duration of 10 seconds were included in the present study. For computational reasons, to train and evaluate the models we used only the first lead and downsampled the signals to 128 Hz. All ECGs with AF or atrial flutter were labeled as 1, and the rest as 0. As regards the prediction task, the dataset is the one described in Section 2.1.3, and it includes more than 350 000 12-lead ECGs recorded at 1 kHz frequency. We excluded by design censored cases with a follow-up shorter than 5 years, reducing to 226 529 signals. Since the prediction task is supposed to be harder than the detection one, in this case we decided to downsample signals to 250 Hz and use all 12 leads for model development and testing. Each ECG was labeled 1 if the corresponding patient developed AF (or atrial flutter) within 5 years from the recording, and 0 otherwise. Notice that the prediction task is approached as a binary classification task.

The architecture used for both tasks is the deep CNN adopted for the study described in Section 2.1.3. The CNN is made of 13 layers with more than 5 million trainable parameters; it takes as input an ECG signal and provides as output the normalized probabilities of belonging to class 0 or 1. To train the models we used the cross-entropy loss function and AdamW optimizer [131]. To avoid overfitting, dropout and data augmentation were applied, the latter implemented adding a zero-mean random Gaussian noise to the training signals. The models were evaluated in terms of Area Under the Receiver Operating Characteristic Curve (AUC). To fit the learning curve, we estimated the model's performance Y (AUC) for different sample sizes n. To do this, we considered a random dataset subsample S that simulated a practical scenario where only a limited amount of data is available. S was set to 1000 (diagnosis) and 10 000 (prediction). For each n in S (detection from 100 to 900 in steps of 50; prediction from 1000 to 8000 in steps of 1000), we randomly sampled without repetition a sample s_n of size n from S, and we trained a model with s_n. Samples in S not included in s_n were used for the evaluation. We repeated this process N times (detection $N = 100$; prediction $N = 50$) for each step, obtaining distributions of AUC values for which we computed means (y_n) and standard errors (SE_n). The data points (n, y_n) were then used to fit the learning curves with an inverse power law function:

$$Y(n) = f(n; a, b, c) = (1 - a) - b \cdot n^c \tag{2}$$

In the fitting procedure, we weighted each data point by $1/SE_n$. NLS and GP were used to fit the learning curves, estimating the parameters a, b, c. To validate the fitted learning curves, we compared the learning curve predictions with the observed AUCs (blue dots in Figure 3) at different target sample sizes. The 95% CI of the predicted AUC were calculated for the GP and NLS as in [155, 156]. Root Mean Square Error (RMSE) and absolute difference were used to compare the test points with the estimated values.

2.2.5 *Results*

Concerning the diagnostic task, via the GP approach after truncating the distribution of the phi parameter to values > 0.01, the curve ($\mu_a = 0.00 \pm 0.00$, $\mu_b = 15.03 \pm 2.60$, $\mu_c = -0.78 \pm 0.03$, $\mu_\varphi = 2.05 \pm 1.98$, $\mu_\sigma = 0.01 \pm 0.01$, $\mu_y = 0.00 \pm 0.00$) provided an adequate prediction of the test points (RMSE= 0.005, Figure 3A). Predicted AUC values of0.95, 0.97, and 0.99 would be reached at $n = 1500$, 3250, and 10 000, respectively. The NLS approach based on Dayimu et al. [156] ($a = 0.00$, $b = 19.28$, $c = -0.83$) also performed well, with RMSE of 0.008 and the test point included in the 95% CI of the predicted AUCs (Figure 3B). Figueroa's NLS implementation [155] ($a = 0.00$, $b = 19.27$,

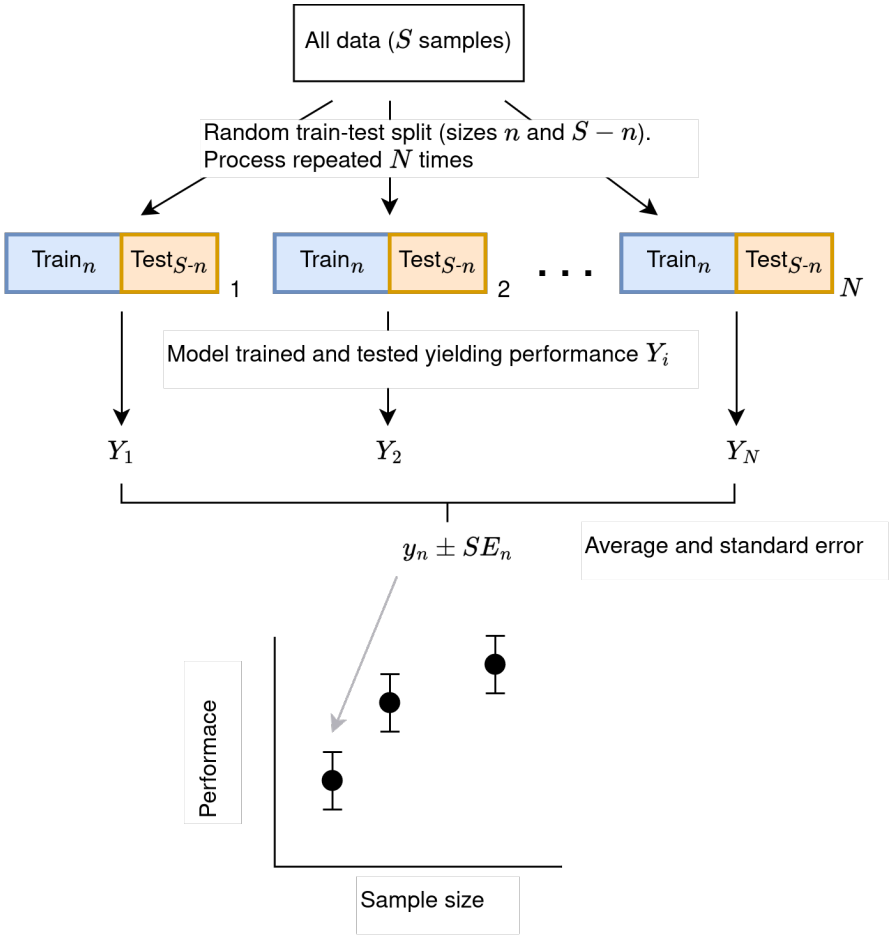


Figure 11: Schema describing the procedure to obtain the learning curve. Figure inspired by [158].

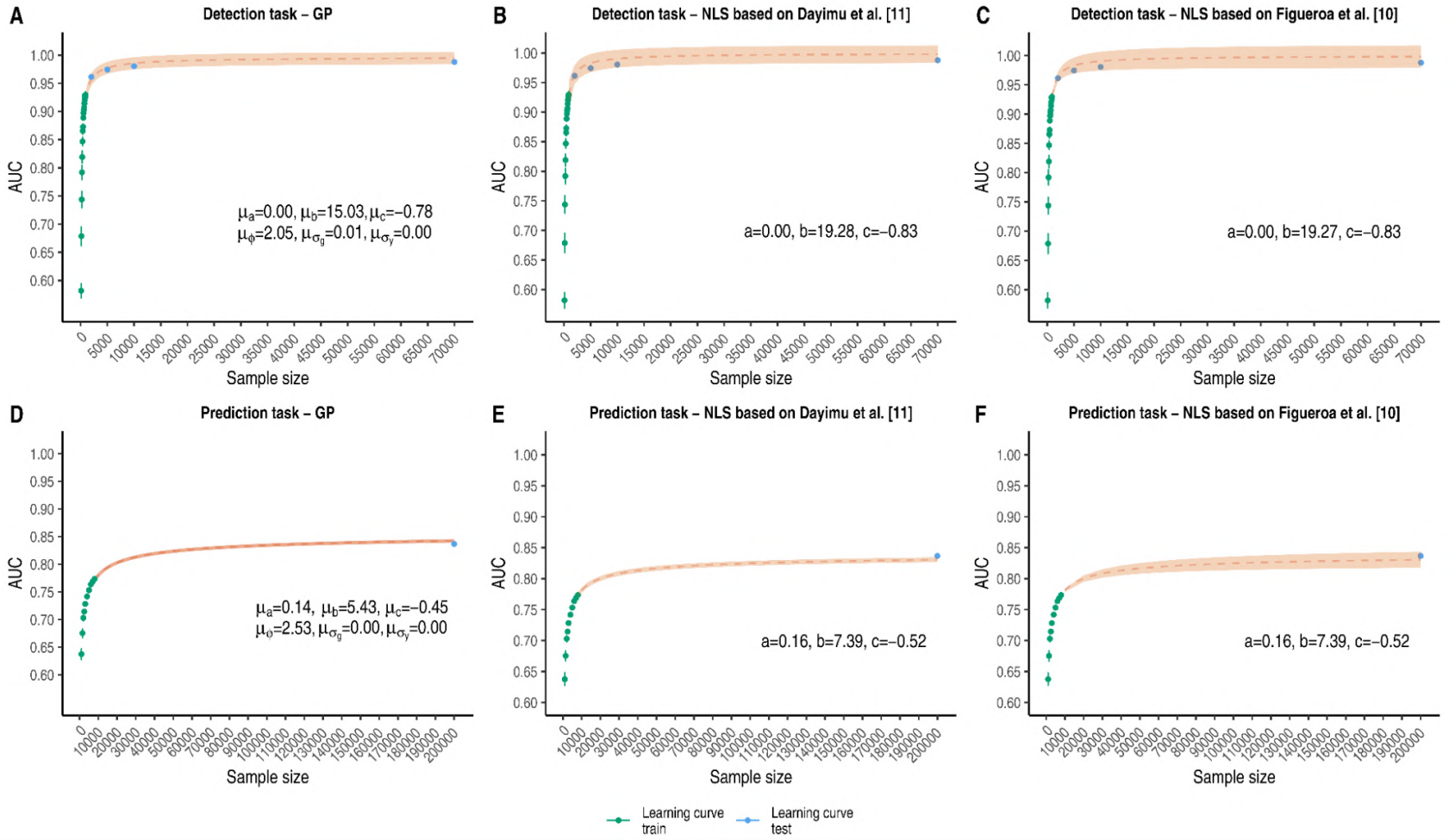


Table 3: Learning curves with 95% CI for the detection (A to C) and the prediction (D to F) tasks based on parameters estimated via GP (A and D) and NLS (B and E based on [156], C and F based on [155]).

$c = -0.83$) provided similar results (RMSE0.008, Figure 3C). Based on these models, predicted AUC values of 0.95, 0.97, and 0.99 would be reached at $n = 1500$, 3000, and 10 000, respectively. Regarding the prediction task, the GP approach ($\mu_a = 0.14 \pm 0.02$, $\mu_b = 5.43 \pm 1.88$, $\mu_c = 0.45 \pm 0.07$, $\mu_\varphi = 2.53 \pm 1.90$, $\mu_\sigma = 0.00 \pm 0.00$, $\mu_y = 0.00 \pm 0.00$) provided an adequate prediction (actual AUC – predicted AUC= 0.006, Figure 3, D). Based on this model, predicted AUC values of 0.82 and 0.84 would be reached at n = 17000 and n=40000, respectively. Similarly, the implementations of the NLS approach based on Dayimu et al. ($a = 0.16$, $b = 7.39$, $c = -0.52$, Figure 3E) and on Figueroa et al. ($a = 0.16$, $b = 7.39$, $c = -0.52$, Figure 3F) performed well (actual AUC – predicted AUC= -0.006). Based on this model, predicted AUC values of 0.80 and 0.82 would be reached at n =20 000 and n =70 000, respectively.

2.2.6 *Conclusions*

We proposed an application of the GP and NLS optimization methods for fitting the learning curve of a CNN algorithm trained to diagnose and predict AF from ECG data. The two methods achieved satisfactory AUC prediction for both tasks. A limitation of the present approach is that it requires a pre-hoc model specification and previously collected and labeled data. The advantages include the flexibility in its application to potentially all types of tasks, performance measures, and architecture models.

2.3 SURVIVAL MULTI-MODAL MODEL FOR AF PREDICTION

2.3.1 *Introduction*

Recent work has highlighted the potential for DL methods to predict Atrial Fibrillation from 12-lead ECGs [11, 12, 138, 144, 161]. However, most of these works approached the AF prediction as a binary classification task: the ECG signal is used to predict the probability that a patient without AF history will develop AF within a certain time window, without including information about the time-to-event or censoring. Khurshid et al. [138] were the first to explicitly incorporate the time until the AF event and missingness due to right censoring in their model, which is important for accurate estimates of absolute risk. To do so, they used *Nnet-survival*, a discrete-time survival model designed for neural networks (Section 2.3.2).

An aspect that was not considered in Khurshid et al.'s work [138] was the presence of competing risks. In the case of AF, death is the primary competing risk, and it should be taken into account when developing a survival prediction model [162]. A method that combines deep learning, survival analysis, and the possibility of handling competing risks is DeepHit (Section 2.3.2).

Combining information from the raw 12-lead ECG signal with EHR data may also improve predictive performance. In this regard, Biton et al. [144] trained a random forest classifier to predict the 5-year risk of AF development using features obtained from different modalities, namely demographics, clinical information, and features extracted from the ECG. The authors showed that the integration of all data sources led to better performance compared to using individual modalities.

The main goal of this work was the development of a survival model to predict new-onset AF from ECGs and EHR data, taking into account death as a competing risk. This was achieved by combining the DeepHit method with a multi-modal DNN able to process both ECG signals and tabular data. As a comparison, we trained other two survival models with the DeepHit approach, but on the single data modalities. To have a further benchmark, we also trained models with the same architectures but in a binary classification setting: in this case, the model predicts the probability that a patient will develop AF within a time window of 5 years. We then compared the predictive accuracy of our model with Cohorts for Aging Research and Genomic Epidemiology (CHARGE-AF) score [110], an AF risk scoring system well-known in clinical literature.

2.3.2 *Background on survival analysis*

Survival analysis is a field of statistics focused on analyzing time-to-event data [163, 164]. Survival time is defined as the time from the beginning of an individual's follow-up to the occurrence of the event of interest. A key concept in survival analysis is *censoring*, which refers to the situation where the exact survival time for some subjects is unknown. Censoring can occur for various reasons: some subjects do not experience the event during the study period, some are lost to follow-up, and others may withdraw from the study due to death or other reasons. Discarding these patients with unknown time-to-event results in a loss of valuable information because it is known that these patients did not experience the event until the study's conclusion. For this reason, various techniques have been developed to model censored data.

Let T and C be the non-negative random variables denoting, respectively, the time to the event of interest and the censoring time. For each individual, we can observe only one of the two. We then define $Y = \min(T, C)$ as the observed time and $\delta = \mathbb{I}(T < C)$ as the event indicator, which is 1 when the event is observed and 0 otherwise. In survival analysis, the label associated with each subject i is the pair (Y_i, δ_i). The random variable T can be described by the following functions:

- *Survival function*. It is the probability that the time to the event of interest is not smaller than a specific time t:

$$S(t) = Pr(T \geq t) \tag{3}$$

  It is a monotonic nonincreasing function, that equals 1 at time 0 and decreases to 0 as time goes to infinity.

- *Hazard function*. It is the instantaneous risk of experiencing the event at time t, given that the individual has survived up to time t:

$$h(t) = \lim_{\Delta \to 0} \frac{Pr(T \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{4}$$

- *Cumulative hazard function*. It is the sum of the hazard up to time t:

$$H(t) = \int_0^t h(s) ds \tag{5}$$

The survival and the hazard functions are linked by the following relationship:

$$S(t) = \exp\left(-H(t)\right) \tag{6}$$

At a population level, the survival function can be estimated using non-parametric estimators, like the Kaplan-Meier (KM) [165] and the Nelson-Aalen [166–168] estimators.

Key tasks in survival analysis include assessing the impact of features on survival time and making predictions. One of the most used models to accomplish these tasks is the Cox Proportional Hazards (PH) model [169], a semi-parametric model that expresses the hazard function as:

$$h(t|\mathbf{x}) = h_0(t)e^{\mathbf{x}\boldsymbol{\beta}} \tag{7}$$

where $\mathbf{x}$ is a vector of covariates (assumed to be fixed at their baseline values in the basic Cox model), $\boldsymbol{\beta}$ represents the regression coefficients, and $h_0(t)$ is the baseline hazard, an unspecified function common to all subjects. This model is termed a *proportional hazard* model because the effect of the covariate is assumed to be independent of time, and, as a consequence, the ratio of hazards of two subjects is a constant. However, this assumption is quite strong and is not always valid.

Competing events

In survival analysis, competing risks occur when there are multiple potential events, and the occurrence of one event prevents the observation of another. For instance, in a study examining time to death due to cardiovascular causes, death from non-cardiovascular causes constitutes a competing event. Competing risks differ from censoring because, in censoring, the event of interest might still occur later, whereas a competing event permanently prevents the primary event from happening. It is crucial to account for competing events when developing risk prediction models, as ignoring them can lead to biased risk estimates [170]. For example, the KM estimator should not be used in the presence of competing risks, as it assumes independent censoring, leading to upwardly biased incidence estimates [162]. Instead, the Cumulative Incidence Function (CIF) should be used, as it is derived from a cause-specific hazard function and provides estimates of the marginal probability of an event in the presence of competing events. When modeling the effect of covariates on time-to-event data in the presence of competing risks, the basic Cox PH model can no longer be used. Alternatives include modeling the effect of covariates on the cause-specific hazard of the outcome [171] or on the CIF [172].

Deep learning for survival analysis

Many machine learning methods have been adapted to handle censored data [173]. These include survival trees [174], random survival forests [175], boosting-based methods [176], deep exponential families [177], support vector machines [178], and Gaussian processes

[179]. Regarding neural networks, the first application of such models to survival analysis dates back to 1995 with the work by Faraggi and Simon [180]. Their approach extended the Cox model by modeling the relationship between covariates and hazard using a feed-forward neural network, generalizing the linear relationship assumed in the original Cox model, eq. (7).

Recently, numerous other Cox-based methods have been proposed [181], and the range of applications is quite wide including tabular data [182], omics data [183], and images [184]. These approaches have improved upon the Cox PH model by relaxing the assumption of a specific relationship between covariates and the hazard function, though they still maintain the PH assumption, i.e. the influence of covariates on hazard remains constant over time.

Another family of DL models for survival analysis is represented by *discrete-time* survival models. In this approach, time is discretized, and the NN models either the discrete hazard rate or the Probability Mass Function (PMF) [185]. The two methods applied in this thesis are *DeepHit* [186] and *Nnet-survival* [187]. DeepHit parametrizes the discrete PMF and accounts for the presence of competing risks. Its loss function combines the log-likelihood of the joint distribution of the first hitting time and the event with a ranking loss for improved discrimination. Nnet-survival (also known as *Logistic-Hazard* [185]), parametrizes the discrete hazard and adopts a loss function which is the negative log-likelihood expressed by the hazard rate. These methods offer greater flexibility than Cox-based methods, as they relax the PH assumption and allow the model to capture the time-dependent influence of inputs on the predicted risk.

2.3.3 *Materials and Methods*

Data

The data source used in this work is the same one used for the studies described in Section 2.1. The only difference is in the time window considered, since in this case we extracted all 12-lead ECGs acquired between February 2, 2007, and December 31, 2022. The same inclusion and exclusion criteria were applied. We obtained, for each patient, demographics, clinical information, and drug prescription/consumption at the exam date by linking the ECG exams with the EHR of the RER of FVG region. The AF events were identified with the same steps described in Section 2.1.3.

In the survival setting, we extracted all available ECGs of the patients without any AF event in the observation period, while for patients that developed AF, we used only the ECGs recorded before the first AF event. We then associated each ECG with a label indicating the type of event the patient underwent (censoring, AF, or death), and the

time to the event. In the binary classification setting, we included by design only censored cases with a minimum follow-up of 5 years. In this case, ECGs were labeled 1 if the corresponding patient developed AF within 5 years, and 0 otherwise.

For each ECG, we extracted from the EHR a set of 62 features. These included demographics, diagnosis, and drug consumption of the patients at the exam date. Furthermore, we decided to include also the wave morphology's features automatically extracted from the ECG signal by the Mortara devices: onset and offset of P and T waves and of the QRS complex, the PR and corrected QT intervals, P, T, QRS axis and the cardiac frequency.

Model development

Regarding the data modality given as input to the model, we considered three different approaches for the AF prediction: ECG signals, tabular data, and the integration of both modalities. For each data modality, we trained a model both in a survival and in a binary setting. Thus, a total of 6 models were trained for this study.

SURVIVAL MODELS In order to build the survival models we used the DeepHit method, which employs a network architecture (sketched in Figure 12) that consists of a single shared sub-network and a number of cause-specific sub-networks (in our case 2, corresponding to AF and death events). The shared sub-network is an encoder that learns a deep latent representation of the input data, and its structure depends on the data modality considered. In the case of ECG signals alone, the encoder's architecture is a Convolutional Neural Network, while for tabular data is a Fully Connected Network (FCN). The integration of both data modalities is obtained by combining the CNN and the FCN with a joint fusion strategy [188], as depicted in Figure 12: each network processes the corresponding data modality and the learned feature representations are joined in a common layer. After the encoder, the latent features are used by each cause-specific sub-network to predict the discrete probability distribution of the corresponding event. Finally, the output layer of the whole network is obtained concatenating the outputs of the cause-specif sub-networks and normalized so that it can be interpreted as the joint probability distribution of the two competing events. More precisely, normalization is done with a single softmax layer that is designed to allow for survival past the maximum follow-up (as described in [185]).

BINARY MODELS As regards the binary prediction task, the networks adopted have the same encoder architectures described above. The difference is that the encoder is now followed by an FCN composed of one hidden layer of l units (where l is the number of latent

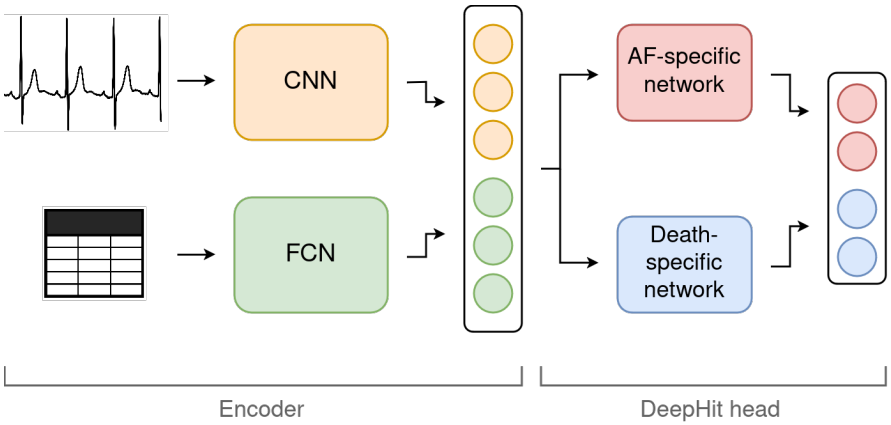


Figure 12: Schema of the deep multi-modal network for survival analysis. For the single-modality networks, only one between the CNN and the FCN constitutes the encoder, while in the classification setting the DeepHit head is substituted by the binary-classification head.

features extracted by the encoder) and an output layer of two units normalized with a softmax function.

NETWORK ARCHITECTURES The CNN's architecture is the one described in Section 2.1.3. The FCN is a multi-layer perceptron with two hidden layers made of $2n$ and n units, where n is the number of tabular features (62 in this study). The cause-specific sub-networks for the DeepHit method are fully connected networks composed of one hidden layer of 10 units, followed by an output layer with a number of neurons equal to the number of discrete time intervals. In all networks, the activation function used is ReLU and dropout is adopted for regularization.

CHARGE-AF SCORE The CHARGE-AF score is a well-validated clinical risk score for the development of AF [110]. The score was derived by estimating a Cox proportional hazards model, trained with the variables age, race, height, weight, systolic and diastolic blood pressure, current smoking, use of antihypertensive medication, diabetes, history of myocardial infarction, and heart failure. 5-year AF risk estimates for CHARGE-AF were computed using the equation $1 - 0.9718412736^{\exp\{(af\_score - 12.58156)\}}$ where af_score is the individual's CHARGE-AF score obtained as a linear combination of the risk factors [110].

Experimental setting

The dataset was split into training, validation and test sets, with a proportion of respectively 7:1:2. In the case of patients with multiple ECGs, we ensured that there was no overlap of patients between the different sets. We also ensured that the positive sample fraction in

each set was as similar as possible to the overall fraction (10%). The validation set was used to evaluate the model during training and to apply early stopping to avoid overfitting. Before feeding them to the network, ECG signals were filtered to remove baseline wander and high-frequency noise with a zero phase second-order infinite impulse response bandpass filter, in the band 0.67–100 Hz[189]. Then, ECGs were also normalized subtracting from each channel its mean and dividing it by its standard deviation. We used a Min-max scaler for tabular data to normalize each feature between 0 and 1. No missing values were present in the tabular dataset.

Since the DeepHit method requires time to be discrete, we had to perform a discretization of the time scale. More precisely, we made an equidistant grid with 50 grid points and we then expressed the time to event of each ECG in this discrete time scale, following the procedure proposed in [185].

Both the survival and the binary classification models were trained via AdamW optimizer, with a batch size of 128 and a learning rate of 10^{-3}. The loss function maximized was the DeepHit loss for the survival model and the binary cross-entropy loss for the binary classifier.

Evaluation metrics

For the survival models, we inspected their discriminative performance computing the area under the time-dependent receiver operating characteristic curve (AUC), with the cumulative/dynamic definition [190]. To account for potential biases, AUC was calculated taking into account the presence of a competing risk event (death) and using inverse probability of censoring weights. To assess calibration, we computed the ICI for competing-risk survival models, as proposed by Austin et al. [191]. ICI is based upon a graphical assessment of calibration: a calibration curve is obtained regressing the cumulative incidence function of the cause-specific outcome of interest on the predicted outcome risk with a Fine-Gray sub-distribution hazard model, and then ICI is computed as the average prediction error weighted by the empirical risk distribution. For the binary classification models, we computed AUC and ICI defined for a binary outcome [122]. Since binary models estimate the risk to develop AF within 5 years, we focused on the predicted 5-year risk also in the survival setting.

Standard errors and corresponding 95% confidence intervals (CI) were estimated using 100-iteration bootstrapping, for both the metrics considered.

2.3.4 *Results*

The final dataset included a total of 350 701 recordings from 128 030 unique patients. See Table 4 for a descriptive snapshot of the population. The median and interquartile age for the recordings was 66

Table 4: Descriptive features of the dataset. For all binary variables we report the percentage of 1s.

		Grouped by AF	
	Overall	0	1
n	351 701	315 348	36 353
Age, median [Q1,Q3]	66 [54, 75]	65 [52, 75]	74 [68, 80]
Gender (males)	50.3	49.9	54.2
Ischemic disease	23.2	22.1	32.8
AMI	14.3	13.8	18.9
TIA/Stroke	4.9	4.5	8.3
COPD	23.9	23.4	28.4
Diabetes	45.2	44.4	51.8
Chronic Heart Failure	8.7	7.6	17.9
Chronic Kidney Disease	14.6	13.5	24.3
Anticoagulants	8.2	7.7	13.0
Antihypertensive	56.6	54.1	78.4
Calcium blockers	18.8	17.5	29.9
Antiarrhythmics	2.7	2.1	8.0
Beta-blockers	30.7	28.9	46.2
Diuretics	12.4	11.6	19.8

[54–75], while the percentage of males was 50.3%. From Table 4 we can also notice that the sample of patients that developed AF was significantly older and with a higher rate of comorbidities with respect to the sample that did not develop AF. The overall median follow-up time was 6.45 years, while it was 4.26 years restricting to individuals that developed AF. In the binary classification setting, we excluded by design censored cases with a follow-up shorter than 5 years, reducing to 226 529 signals for 95 823 unique patients.

In Table 5 we report 5-year AUC and ICI values (and corresponding 95% CI) for the six trained models. The first thing that we can notice is that the survival model trained combining ECG and EHR data shows good performance, with an AUC of 0.845 and an ICI of 0.010. To visually inspect the model's output, we plot in Figure 13 the observed cumulative incidence function for the AF event, stratified by the 5-year risk predicted by the model. Looking at the figure, it is possible to see that the incidence curves differ significantly, and in particular that patients with higher predicted risk are associated with higher AF incidence at 5 years.

As regards the comparison with the other approaches, we can notice from Table 5 that, from a discrimination point of view, the sur-

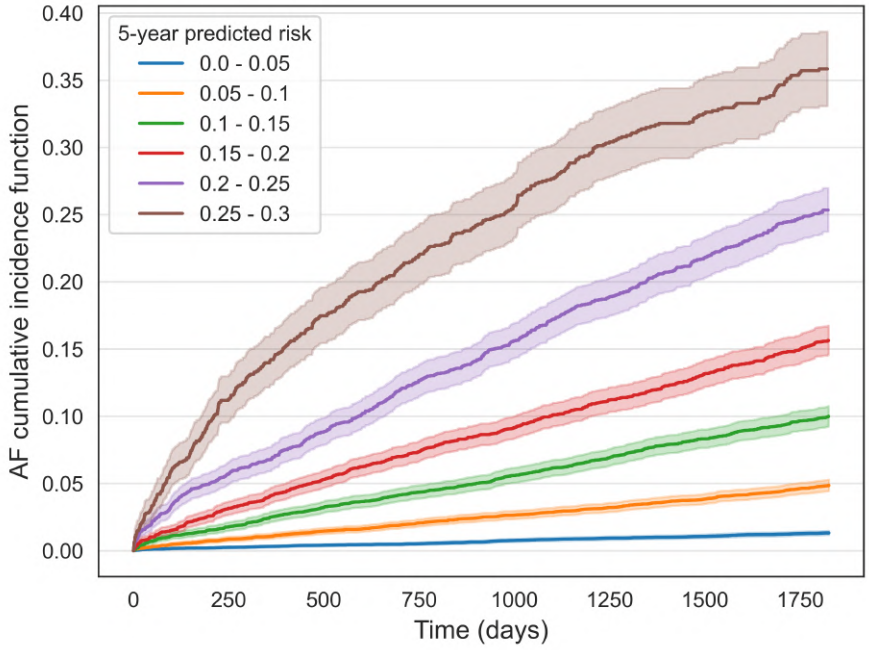


Figure 13: Observed cumulative incidence function for the AF event, stratified by the 5-year AF risk predicted by the DeepHit model trained with both ECGs and EHR data. Cumulative incidence is estimated with the Aalen-Johansen estimator. Shaded areas around curves indicate 95% confidence intervals.

vival and binary classification models behave in a very similar way, while the most influential factor on discrimination is the data modality used to train the model. In particular, the best-performing models in terms of AUCs are the ones that integrate ECG information and clinical tabular data, with AUCs of 0.845 and 0.842 for the survival and the binary models, respectively. The use of ECG signals alone leads to models with an AUC of approximately 0.82, while the weakest discrimination is obtained when only tabular data are used to train the algorithm (AUC around 0.81). Moving to calibration, we can see that all the models are well calibrated in general, with the maximum ICI value of 0.024 obtained for the binary model trained with tabular data alone. The second worst calibrated model is again the one trained only with tabular data, but in the survival setting, with ICI equal to 0.017. For the rest of the models, the calibration performance is similar.

We compared the survival model trained with both ECGs and tabular data to the CHARGE-AF clinical score, as reported in Table 6. For this comparison, we restricted to the subset of the test set for which we had no missing values in the predictors of CHARGE-AF. The CHARGE-AF score demonstrated moderate discrimination and calibration, with AUC and ICI of 0.744 and 0.036, respectively. In the same sample, the DeepHit model showed consistently better performance

Table 5: Models Performance for AF prediction in Test Sets

Model	**5-year AUC**	**5-year ICI**
DeepHit (ECG + tabular)	0.845 (0.839, 0.851)	0.010 (0.008, 0.011)
DeepHit (ECG)	0.820 (0.812, 0.826)	0.005 (0.004, 0.007)
DeepHit (tabular)	0.810 (0.804, 0.816)	0.017 (0.015, 0.019)
Binary (ECG + tabular)	0.842 (0.835, 0.849)	0.006 (0.004, 0.009)
Binary (ECG)	0.821 (0.814, 0.827)	0.013 (0.010, 0.016)
Binary (tabular)	0.812 (0.806, 0.819)	0.024 (0.021, 0.027)

Table 6: Comparison between model performance and CHARGE-AF clinical score, on the subset of patients that do not have missing values in CHARGE-AF predictors.

Model	**5-year AUC**	**5-year ICI**
DeepHit (ECG + tabular)	0.811 (0.800, 0.824)	0.009 (0.007, 0.013)
CHARGE-AF	0.744 (0.735, 0.754)	0.036 (0.032, 0.043)

both in terms of discrimination (AUC of 0.811) and calibration (ICI 0.009). Notice that DeepHit discriminative performances decreased significantly when compared to the results on the entire test set (Table 5). This may probably be related to different sample characteristics, since, for example, the CHARGE-AF sample is characterized by a higher risk of developing AF (the fraction of ECGs associated with a future AF event is approximately 12% in the CHARGE-AF sample, against the 10% of the entire dataset).

2.3.5 *Discussion*

In this study, we developed a deep learning model for the prediction of incident AF using 12-lead ECG signals and EHR data, explicitly incorporating in the model time to the event and censoring information. The prediction of AF is a critical medical task due to its high prevalence in the elderly population and the importance of an early diagnosis that could prevent stroke or other fatal outcomes.

The model was trained on roughly 240 000 ECGs and it showed good performance on the test set, both in terms of discrimination (AUC 0.845) and calibration (ICI 0.010). When compared to the models trained on single-modality data (ECG or EHR), the complete model showed better discriminative performance and a similar calibration. Taking into account that the higher the model complexity is the lower calibration is expected, this is a very interesting result. Moreover, it confirms the intuitive idea that the fusion of data from multiple

modalities can improve the model's predictive power, probably due to the fact that each modality contains complementary information that the model smartly integrates for the prediction. We also compared our model to the CHARGE-AF risk score. What we have found is that our model consistently outperforms CHARGE-AF in terms of discrimination and calibration.

Our results seem to indicate that, given the data modality used for training, survival and binary approaches lead to a similar predictive performance. However, there is more than one reason why it should be preferred to work in a survival setting. First of all, the use of a survival model makes it possible to leverage information about all patients. Indeed, in the binary classification task, censored patients with a follow-up shorter than the study time window should be discarded by design. Conversely, when training a survival model all censored individuals can be used, and they contribute to the loss function only at time bins occurring before censoring. We expect this aspect to be relevant in particular in case of smaller sample sizes. Second, a survival model gives the possibility to predict risk at different time distances, while a binary classification model could consider only a fixed time window at a time. Last, but not least, the survival setting that we used takes into account also the survival time of the competing event, and in general long-term cohort studies this aspect is quite relevant (for increasing incidence of the competing event). To the best of our knowledge, our model is the first deep learning model, in the context of AF prediction from ECGs, to account for competing risks. This aspect is fundamental since failing to consider competing events during model development can lead to an overestimation of the predicted risk [126, 170]. In our cohort, the percentage of patients with death events was approximately 17%, and thus we decided to account for it in our model.

The prediction of AF from ECG signals has been investigated by recent literature. Raghunath et al. [12] developed a deep-learning model to predict incident AF using >1 million 12-lead ECGs, demonstrating good discrimination at 1 year (AUC of 0.83, 95% CI 0.83–0.84). The authors also showed that including age and sex in their model slightly improved discrimination (AUC 0.85, 95% CI 0.84–0.85). Although the sample size and the time-frame prediction period considered were clearly different from ours, the performance is comparable with our findings (Table 5). However, no measures of calibration were reported in the study, a fundamental metric for clinical risk prediction models [116]. Khurshid et al. [138] trained a CNN to infer 5-year incident AF risk using 12-lead ECGs, and were the first to explicitly incorporate survival time and censoring in this context. An aspect that authors did not account for was death as a competing risk, given the low death rates within the time window of interest (4.6% in the internal test set). In our cohort, the death rate at 5 years was much higher

(12.2%) and could not be ignored. The authors obtained the best predictive performance (AUC 0.838 [95% CI, 0.807 to 0.869], ICI 0.012) fitting a proportional hazard model composed of the 5-year risk predicted by the CNN and the CHARGE-AF risk score. They also assessed model performance on 2 external test sets, an aspect which is lacking in our study. A recent study that combined information derived from the raw 12-lead ECG with clinical information to predict AF development is the one by Biton et al. [144]. In their work, authors trained a random forest classifier using EHR variables, ECG-engineered features and features extracted from the ECG with a previously trained deep learning network. They obtained a very accurate model, with an AUC of 0.909 (0.903, 0.914), indicating the great potential of integrating data from different sources. However, it should be noticed that their data-fusion approach is different from ours, since in our case the latent feature representation learning is not separated from the prediction model, instead they are processed by the same network. This should make it possible to learn better feature representations for the different modalities [188]. Compared to our study, Biton et al. worked with a much larger sample size (more than 1 M recordings) and with a considerably smaller mean follow-up (1.25 years, against 6.45 years in our study population). Since we expect that the distance to the event plays an important role in the prediction of AF development, this is a detail that should not be ignored. Model calibration was not assessed in this study and time to event was not integrated in the model.

Our study has some limitations. First of all, we could not validate our results in an external validation cohort, which represents a critical step in the development of machine learning models in medicine to assess the generalizability of the prediction algorithm [137]. Second, we did not perform any explainability analysis of our model. Indeed, it would be of interest to understand which parts of the ECG and which clinical features have the greatest influence on model-predicted risk estimates. This is an aspect that we would like to explore in future works. Another aspect where there is room for improvement is the way we integrate data from different sources. Indeed, the choice we made is straightforward and maybe it is not the best one to learn a joint embedding space of the two data modalities. In this regard, we expect that self-supervised learning, the machine learning paradigm in which unlabeled data are processed to obtain useful latent representations that can improve downstream learning tasks [192], could help. Recent findings have shown that self-supervised representation learning can significantly enhance model performance when applied to both ECG signals [106, 193] and tabular data [194]. Thus, we would like to explore this technique in a multi-modal setting [195].

2.3.6 *Conclusion*

We are the first to develop a survival deep learning model for AF prediction that accounts for death as a competing risk. We showed that integrating data from different modalities (ECG signals and EHR tabular data) improved model performance with respect to models trained on single modalities, in line with previous findings.

3

CLUSTERING FOR PATIENT PHENOTYPING

This chapter focuses on applications of clustering methods in two different contexts: DCM patient phenotyping and personalized treatment effect characterization. The first section is dedicated to a brief introduction of the clustering methods used in the mentioned works, which are described in the following two sections.

3.1 BACKGROUND

Clustering is a widely used technique in unsupervised learning that partitions a dataset into distinct groups, or *clusters* [196]. The goal is to group similar observations while ensuring that observations in different clusters are distinct. Clustering helps uncover underlying patterns and structures in unlabeled data, with each cluster potentially representing a unique concept or category.

There are numerous clustering algorithms available. In the following work, we applied two of the most well-known methods: K-means [197] and hierarchical clustering [198]. K-means aims to minimize the intracluster distance while maximizing the intercluster distance. Given a dataset $\{\mathbf{x}_1 \dots \mathbf{x}_N\}$, K-means algorithm partitions the dataset into $p \leq N$ clusters $C_1 \dots C_p$ minimizing the function [34]

$$E = \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in C_i} \|\boldsymbol{x} - \boldsymbol{\mu}_i\|_2^2, \tag{8}$$

where $\boldsymbol{\mu}_i$ is the *centroid* of cluster C_i, i.e. the mean vector

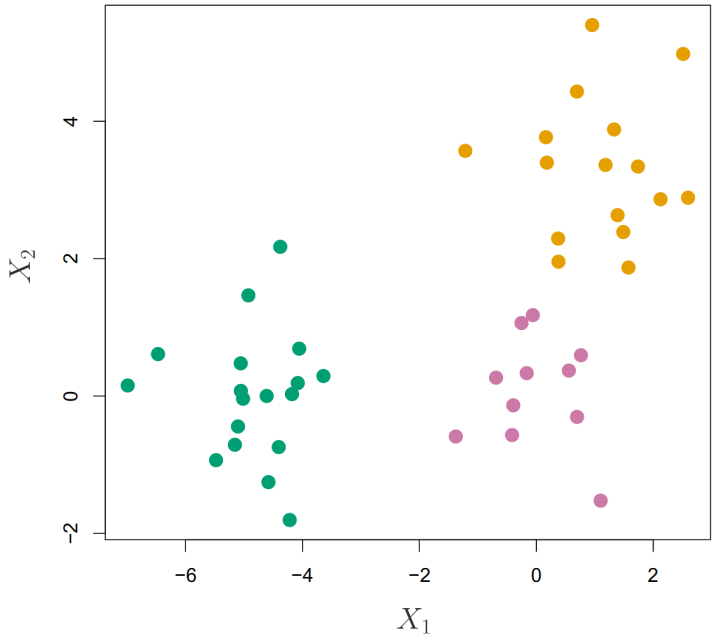
$$\boldsymbol{\mu}_i = \frac{1}{|C_i|} \sum_{\boldsymbol{x} \in C_i} \boldsymbol{x}. \tag{9}$$

Intuitively, minimizing (8) is equivalent to finding the partition such that the total within-cluster variation, summed over all p clusters, is as small as possible. Notice that this problem is NP-hard and cannot be solved exactly. Therefore, K-means finds an approximate solution by an iterative algorithm:

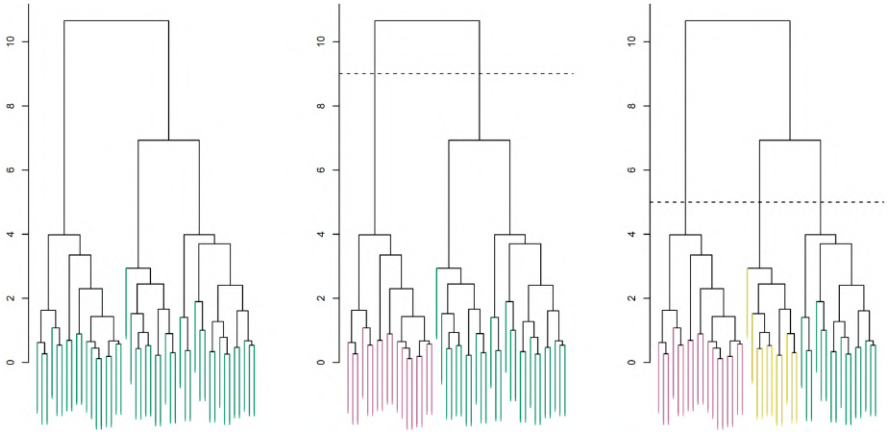
1. Randomly assign each observation to one of p clusters as an initial grouping.
2. Repeat until cluster assignments remain unchanged:
   a) Calculate the centroid of each cluster
   b) Reassign each observation to the cluster with the nearest centroid, based on Euclidean distance.

Note that the desired number of clusters must be decided in advance for K-means. In contrast, hierarchical clustering does not require prespecifying the number of clusters. Instead, it creates a hierarchy of clusters that can be visualized through a tree-like structure called a dendrogram (Figure 14). This approach can follow either a bottom-up or top-down strategy. The bottom-up approach, also known as *agglomerative* clustering, starts by treating each sample as a separate cluster. At each iteration, the two closest clusters are merged to form a new cluster. To determine which clusters to merge, a distance measure between clusters is needed, with several options available. The simplest approach is to compute all pairwise dissimilarities between the elements in two clusters and consider the minimum, maximum, or average value as the cluster distance. These methods are referred to as *single-linkage*, *complete-linkage*, and *average-linkage*, respectively. Another popular method, known as Ward's linkage, calculates the distance between two clusters as the difference between the total within-cluster sum of squares for the two clusters separately and the within-cluster sum of squares after merging them. In this case, the algorithm is equivalent to minimizing the total within-cluster variance, making it the hierarchical analog of K-means.

K-means and hierarchical clustering, together with other clustering approaches, have been extended to Functional Data Analysis (FDA), the branch of statistics that analyzes data that are functions or curves rather than discrete data points [199, 200]. These methods enable the identification of heterogeneous morphological patterns in continuous functions, and a practical example is explored in Section 3.3.



(a)



(b)

Figure 14: *(a)* Simulated observations in a two-dimensional space. *(b)* **Left**: Example of dendrogram obtained from hierarchical clustering the data from panel *(a)* with complete linkage and Euclidean distance. **Center**: the dendrogram from the left-hand panel, cut at a height of nine (dashed line). This cut results in two distinct clusters, shown in different colors. **Right**: the dendrogram from the left-hand panel, cut at a height of five, resulting in three distinct clusters [196].

3.2 DILATED CARDIOMYOPATHY PHENOTYPING

This work is based on the study:

- Ilaria Gandin, Maria Perotto, Alessia Paldino, **Giovanni Baj** et al. "Clustering in Dilated Cardiomyopathy at Initial Evaluation: An Effective Tool for Clinical Stratification". Submitted to *European Journal of Heart Failure*.

3.2.1 *Introduction*

Dilated Cardiomyopathy (DCM) is a heterogeneous disease in terms of clinical presentation and genetic background. Despite multiple phenotypic and genetic findings have been suggested as possible prognostic features [201–203] of Heart Failure and Malignant Ventricular Arrhythmias (MVA), no definite prognostic risk score has yet been validated in this field [204]. AI and ML recently helped in the characterization of distinct phenogroups within the DCM spectrum, each associated with a unique clinical presentation, distinct underlying aetiology, and outcomes [73, 205]. However, this approach required the collection of advanced information, including second and third-level exams, such as CMR data and even endomyocardial biopsy. Similar results were obtained from cardiac transcriptome of patients affected by DCM [206]. While these studies demonstrate common features shared by many DCMs, the starting data used for these unsupervised clustering are often not available upon first medical contact, not always collected during follow-up, and therefore not applicable in all cardiac centres. ECG remains the simplest, most reliable, and lowest-cost technology with excellent diagnostic and prognostic functions. To prove ECG importance, in the "Madrid Genotype Score", which estimates the probability of receiving a positive genetic test in DCM, 2 out of 5 criteria are ECG parameters (low voltage on the electrocardiogram and absence of Left Bundle Branch Block (LBBB) [207]). Leveraging the potential of ML in stratifying patient prognosis in DCM only with data that can be collected upon first medical contact, the aims of our study were: (I) to apply unsupervised clustering to patients affected by DCM using clinical information available at first medical contact; (II) to identify clustered patients based on phenotypic similarities; (III) to analyse these clusters in terms of genetic background and outcomes.

3.2.2 *Methods*

Study cohort and study design

This observational retrospective study includes a primary cohort of patients with DCM who underwent genetic testing, recruited from the Familial Cardiomyopathy Registry of Trieste [201, 208]. DCM was

defined as left ventricular dilatation and impaired ejection fraction (LVEF$<$ 50%), following a thorough exclusion of secondary causative etiologies [73, 209]. Of note, Non-Dilated Left Ventricular Cardiomyopathy (NDLVC) cases with only systolic dysfunction (i.e. LVEF $<$ 50% at enrolment) were included. All patients underwent a baseline evaluation upon study enrolment. Collected information included demographic and clinical data (inclusive of HF symptoms and NYHA functional class), information on family history of cardiomyopathy and SCD, results of 12-lead ECG and 24-hours Holter ECG monitoring and echocardiographic assessment. Transthoracic echocardiogram with biventricular dimensions and systolic function was assessed according to international guidelines [210]. Regarding ECG, a wide collection of features was recorded (*Supplemental table S1*). Results of genetic testing, endomyocardial biopsy, CMR, and cardiac exercise stress testing information were not included in the clustering dataset as deemed often unavailable upon first cardiological evaluation. The study included a validation cohort of DCM patients enrolled in the CARITMO (CARdiopatie ARITMOgene) registry of the Casilino Hospital in Rome (Italy). The study was approved by the ethical committee (CEUR N.O. 43/2009, Em 06/22).

Genetic testing

All enrollees were screened for genetic variants associated to DCM by Next Generation sequencing of multigene panels, as previously reported [201, 208, 211]. Variants were classified as pathogenic or likely pathogenic (P/LP) according to the American College of Medical Genetics and Genomics criteria [212].

Endpoints

There were three composite endpoints: 1) all-cause mortality or Heart Transplantation (D/HT); 2) severe arrhythmic events (SCD/MVA); and 3) heart failure-related events (HF death/HT/ left ventricular assist device implantation). SCD included witnessed SCD with or without documented Ventricular Fibrillation (VF), death within 1 hour of acute symptoms, or nocturnal death with no antecedent history of immediate worsening symptoms. MVA was defined as VF, sustained ventricular tachycardia (lasting $>$ 30 seconds or with hemodynamic instability), and appropriate ICD interventions (shock or antitachycardia pacing on VF or sustained ventricular tachycardia).

Statistical analysis

CLUSTERING Given the framework of the study, which is based on a data-driven approach, the principle followed in the data collection for clustering analysis was to retain as much information as possible. This resulted in a large set of covariates that encompass the

outcome of the evaluations, with high levels of correlations. Covariates were retained if the missing rate was lower than 10%. Binary variables with lower frequency class $< 2\%$ were also excluded. The analysis presented two major challenges: a large number of variables and their mixed-data type (presence of both numerical and categorical variables). For this reason, a two-step approach was followed. First, a dimensionality reduction technique called Principal Components Analysis for mixed data (PCAmix), which extends the principal component analysis to mixed data, was applied (see Supplemental Materials B). Secondly, a clustering method was used on the reduced dataset formed by 11 principal components. More specifically, an agglomerative hierarchical clustering algorithm was applied. The optimal number of clusters was identified using the average silhouette criteria [213]. Genetic information was not part of the input features of the cluster analysis since the focus of the present clustering was a phenotypic characterization with elements the clinician has available upon first medical evaluation. Instead, the presence of P/LP variants was investigated a posteriori in relationship with cluster partition to understand possible differences between clusters in terms of disease aetiology. For this purpose, genes DSP, PKP2, FLNC, and LMNA were grouped as "arrhythmic genes" and all the remaining as "non-arrhythmic genes".

SIMPLIFIED CLUSTERING AND VALIDATION To provide a clustering rule that can be easily applied in clinical practice, and to the external cohort of the present study, a simplified version of the clustering model, involving only a reasonable number of variables, was obtained using a penalized logistic regression. A LASSO model (a standard technique for variable selection problems [214]) was estimated in the study cohort with all the input variables involved in the clustering as predictors, and the clusters assignment as outcome. The penalty parameter was selected such that only three predictors were not shrunk to zero. The performance of this simplified model was evaluated in terms of ROC-AUC through a 10-fold cross-validation. Based on the continuous score of the simplified model (ranging from 0 to 1), individuals were assigned to the second cluster if exceeding the value 0.23, corresponding to Youden's cut-point.

GROUP COMPARISON Continuous variables were compared between groups using the t-test or the nonparametric Mann-Whitney U test as appropriate. Discrete variables were analyzed using the chi-square or Fisher's exact test. For the D/HT endpoint, Kaplan-Meier survival curves were estimated and compared between groups with the log-rank test. For the SCD/MVA endpoint, cumulative incidence curves at 20 years of maximum follow-up were obtained considering D/HT and HF/HT competing events and compared with the Gray's test.

The same approach was followed for the HF/HT endpoint, for which SCD/MVA and D/HT are competing events. To include the effect of known risk factors, multivariable cause-specific Cox models were obtained and nested models were compared using the likelihood ratio test and Harrell's C- index. Martingale residual plots were visually inspected to assess linearity. Statistical analyses were performed in R (version 4.4.1) using the packages `PCAmixdata`, `cluster`, `survival`, and `cmprisk`.

3.2.3 *Results*

Derivation cohort and clustering

The derivation cohort consisted of 409 patients affected by DCM. The majority were men (291, 71%) and the mean age at the time of recruitment was 46 years (SD= 14). Table 7 (first column) reports the main demographic and clinical characteristics at the baseline of the derivation cohort. 102 input features were initially considered for the cluster analysis. Following the PCAmix strategy, we obtained a dataset with reduced dimensionality preserving the 47% of the variance, which were fed to the hierarchical clustering algorithm. This clustering was able to recognize two main DCM clusters: a first cluster (CL1) of 334 individuals (82%) and a second cluster (CL2) of 75 individuals (18%). Comparing characteristics at baseline, subjects of CL2 showed a less arrhythmic profile (fewer Premature Ventricular Contraction (PVC)/24h and Non-Sustained Ventricular Tachycardia (NSVT)) and a worse LVEF, compared to CL1 (Table 8). Features that most differentiated the two groups were mainly ECG parameters (Figures 15 and 16). Compared to CL1, group CL2 had a higher prevalence of true LBBB (1% vs 76%, $p < 0.001$), intrinsicoid deflection in V5 or V6 (7% vs 87%, $p < 0.001$), and Left Ventricular Hypertrophy (LVH) by Cornell criteria (16% vs 75%, $p < 0.001$), together with higher QRS duration in V1 (median value 100 vs 160 ms, $p < 0.001$), higher QRS duration in V6 (median value 105 vs 165 ms, $p < 0.001$), higher S wave voltage (in V1 or V2, median value 14 vs 30 mm, $p < 0.001$), and higher S wave duration in V2 (S nadir-to-end, median value 40 vs 80 ms $p < 0.001$).

Clinical outcomes in phenotypic groups

During a median follow-up of 100 months (IQR 51-185), 66 individuals met the primary endpoint (death/HT), 92 SCD/MVA events, and 46 HF events were recorded. For the primary endpoint and the HF-related endpoint, there were no differences between CL1 and CL2 (Figures 36 and 37 in the Supplementary material). Instead, a lower risk for SCD/MVA events was observed in CL2 ($p = 0.006$, Figure 17) with an associated hazard ratio of 0.29 (95% CI 0.13- 0.67). When con-

Table 7: Baseline characteristics of the study cohort and the external validation cohort. Fewer data are available for the external validation cohort as they were collected before the model was designed on the primary cohort. CMP: Cardiomyopathy

	Cohort N=409	**Val. cohort N=160**	**P-value**
Men, N (%)	291 (71)	108 (68)	0.4
Age, Mean (SD)	46 (14)	54 (13)	<0.001
European, N (%)	408 (99)	160 (100)	0.9
Genetic positive P/LP variants, N (%)	169 (41)	33 (39)	0.8
Arrhythmic gene P/LP variants, N (%)	53 (31)	14 (42)	0.3
Syncope, N (%)	34 (8)		
NYHA, N (%)			
1	207 (51)		
2	121 (30)		
3	59 (14)		
4	22 (5)		
Fam. hist. of CMP N (%)	139 (34)		
Fam. hist. of SCD N (%)	62 (15)		
LVEF %, Mean (SD)	35 (12)		
AF, N (%)	14 (3)		

Table 8: Comparison of baseline characteristics between CL1 and CL2.

	CL1 N=334	**CL2 N=75**	**P-value**
DCM clinical characteristic			
Men, N (%)	251 (75)	40 (53)	<0.001
Age, Mean (SD)	45 (14)	50 (10)	0.001
Genetic positive, N (%)	158 (47)	11 (15)	<0.001
PVC/24 h, N (%)	68 (20)	4 (5)	0.002
NSVT, N (%)	142 (43%)	21 (28%)	0.020
LVEF, Median (IQR)	35 (26-45)	30 (23-38)	0.005
LVEF, <35%	165 (49%)	51 (68%)	0.004
LV diastolic diameter, Median (IQR)	62 (57-68)	62 (58-72)	0.5
AF, N (%)	14 (4.2%)	0 (0%)	0.083
Cluster distinctive features			
True LBBB, N (%)	2 (1)	57 (76)	<0.001
Intrinsicoid deflection V5-V6, N (%)	25 (7)	65 (87)	<0.001
LVH by Cornell criteria, N (%)	52 (16)	56 (75)	<0.001
QRS duration V1, Median (IQR)	100 (90, 110)	160 (142, 170)	<0.001
QRS duration V6, Median (IQR)	100 (90- 110)	160 (142- 170)	<0.001
S nadir to end (ms), Median (IQR)	40 (40, 40)	80 (60, 90)	<0.001
S amplitude V2 (mm), Median (IQR)	14 (10, 20)	30 (25, 38)	<0.001

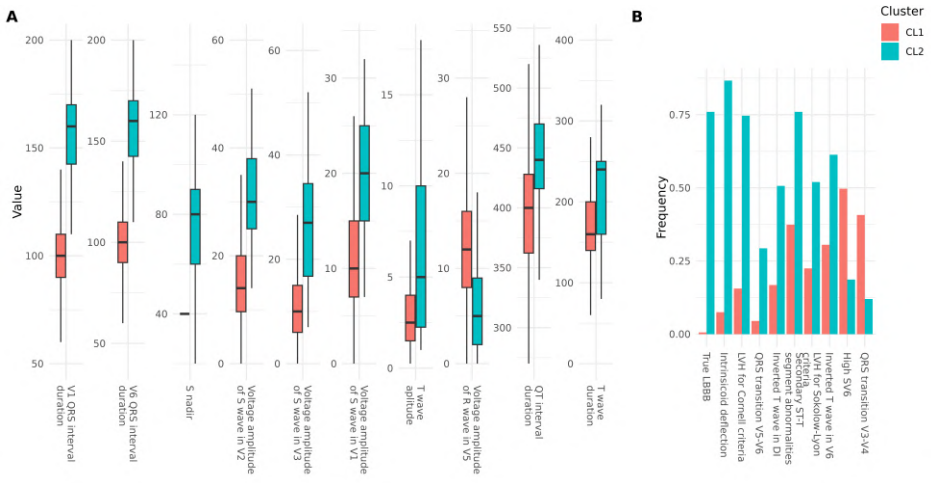


Figure 15: First 10 continuous variables ordered by mean difference after standardization (A) and the first 10 binary variables ordered by standardized chi-square residuals (B) are shown to provide a description of the clusters' composition.

Table 9: Multivariable analysis for SCD/MVA events including cluster 2 (CL2).

Characteristic	**HR (95% CI)**	**p-value**
Sex	1.56 (0.90, 2.73)	0.12
Age	1.02 (1.00, 1.04)	0.069
Family history of SCD	1.33 (0.73, 2.40)	0.3
T-wave inversion	1.54 (0.89, 2.68)	0.12
Syncope	4.14 (2.35, 7.30)	< 0.001
NYHA 3-4	1.54 (0.87, 2.70)	0.14
LVEF	0.98 (0.96, 1.01)	0.2
CL2	0.20 (0.08, 0.48)	< 0.001

sidering the effect of sex, age, family history of SCD, lower lateral T waves, syncope, NYHA 3 or 4, LVEF, the inclusion of CL2 information had a significant impact (adjusted hazard ratio 0.20 (95% CI= [0.08, 0.48, $p < 0.001$; increased C-index from 0.69 to 0.74; see Table 9). Given the higher prevalence of true LBBB in CL2, the role of Cardiac Resynchronization Therapy (CRT) was investigated to determine whether CRT had a role in CL2 prognosis. Considering a prognostic model including CRT presence, the inclusion of CL2 information improved the model ($p = 0.0002$) and increased C-index from 0.68 to 0.74.

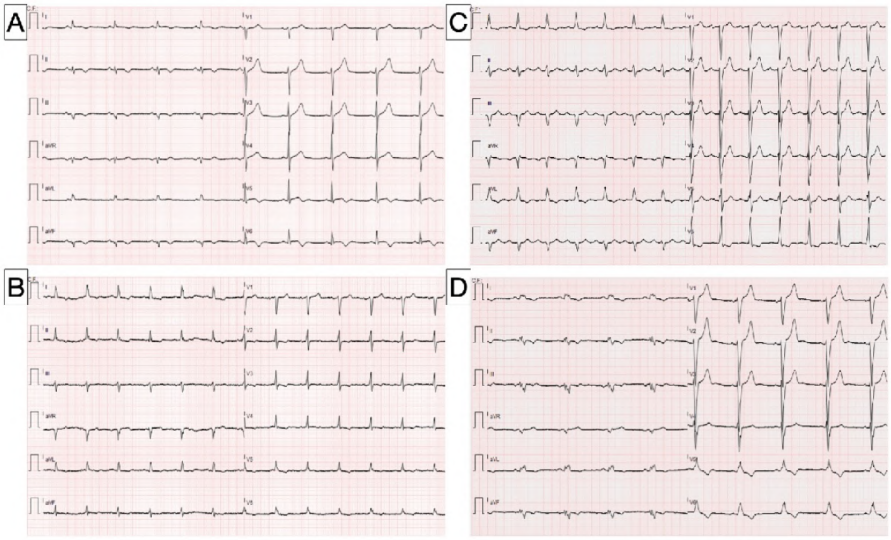


Figure 16: Examples of electrocardiograms of patients belonging to cluster 1 (panels A and B) and cluster 2 (panels C and D).

Genetic yield in phenotypic groups

The relationship between the genetic aetiology of the disease and the two clusters was analyzed a posteriori. A lower yield of P/LP variants was found in CL2 compared to CL1 (15% vs 47%, $p < 0.001$). Given the association found between CL2 and a lower risk for SCD/MVA events, the rate of P/LP variants in arrhythmic genes (DSP, PKP2, LMNA) was compared: CL1 included 20 cases (6%) and CL2 none (0%, $p = 0.033$).

Validation in external cohort

Using a penalized logistic model, we obtained a simpler clustering rule that was able to classify individuals in CL1 and CL2 using only 3 variables, maintaining a high accuracy (AUC=0.991, Standard Error (SE)=0.005). Variables involved were QRS duration in V6, true LBBB, and intrinsicoid deflection > 50 ms (see Table S2 in Supplemental for details). This simplified model was applied to the validation cohort, which included 160 patients, 108 men (68%) with a mean age at baseline of 54 years (SD=13) (see Table 7, second column). In the validation cohort, 126 individuals (79%) were assigned to CL1 and 34 individuals (21%) to CL2. During a median follow-up of 50 months (IQR 25-84), 3 patients died, and 21 patients experienced a SCD/MVA event (no HF events were recorded). There was no significant association between clusters and D/HT events ($p = 0.3$). Instead, CL2 was associated with a lower risk of SCD/MVAs ($p = 0.017$, Figure 18). More specifically, none of the individuals belonging to CL2 experienced SCD/MVA events.

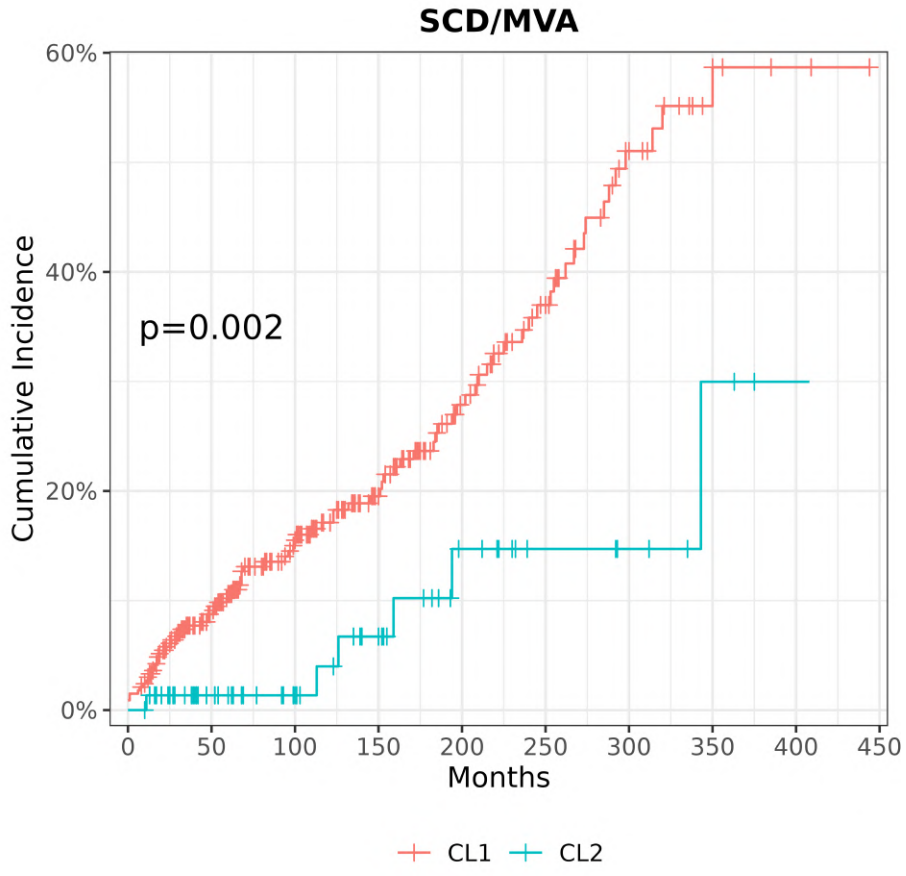


Figure 17: Comparison of cumulative incidence curves for SCD/MVA events between CL1 and CL2 in the study cohort.

3.2.4 *Discussion*

With this study, we aimed to explore the potential of clinical information collected at the very first evaluation (ECG, echocardiographic, Holter ECG and clinical data) in recognizing distinct high-risk subgroups of patients with a diagnosis of DCM or NDLVC with systolic dysfunction by the application of an unsupervised clustering methodology. In this study, two clusters were identified, which differed mainly in terms of ECG presentation. The two groups showed an interesting antithetical genetic background and arrhythmic outcomes during follow-up. These clusters were able to predict arrhythmic outcomes even when put together with commonly recognized risk factors for MVA, such as LVEF or family history of SCD. For a future clinical application, we were able to simplify clustering variables to only three of them, retaining an effective ability to distinguish clusters and showing reproducibility in an external validation cohort from another center. This is the first study showing how ECG data, easily available at first evaluation, can effectively cluster patients affected by DCM. This is particularly of interest when considered that many other data

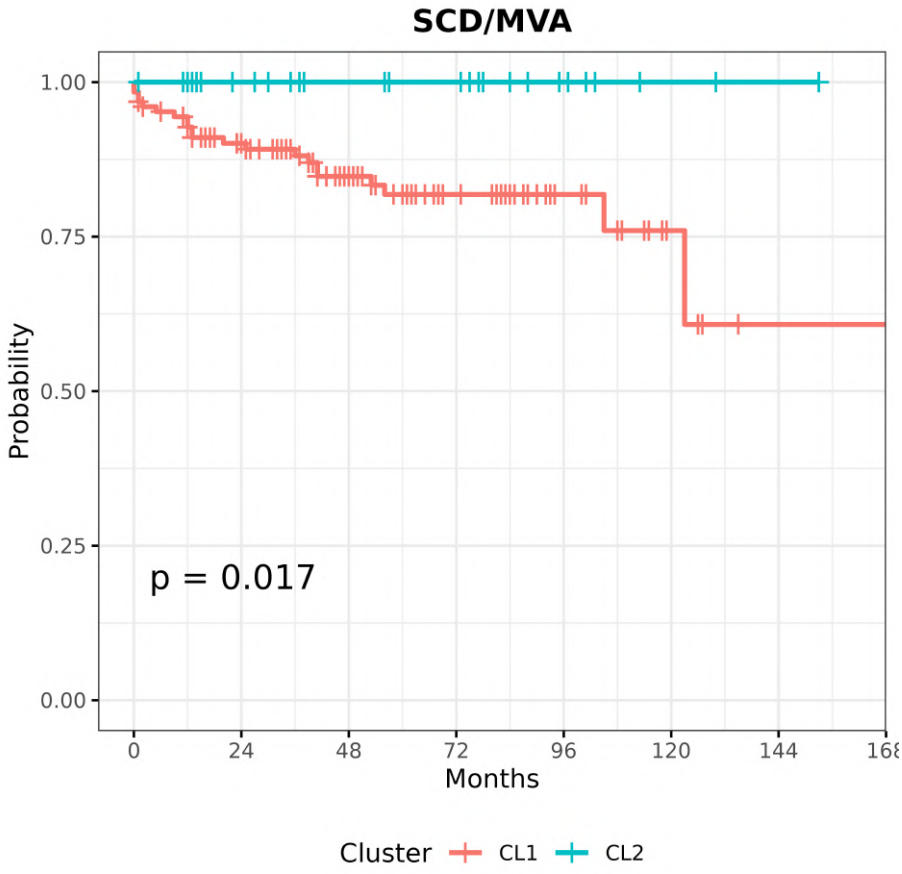


Figure 18: Comparison of Kaplan-Meier curves for SCD/MVA events between CL1 and CL2 in the validation cohort.

were fed into the algorithm, such as echocardiographic data, but a cardiomyopathy-oriented ECG interpretation was enough to distinguish the two clusters, to predict patient prognosis, and to show an association with genetic background. If CL1 exhibited high variability in terms of clinical findings, genetic background, and outcomes, our clustering appears to have succeeded in identifying mostly CL2, which has a significantly and homogeneously better prognosis. The CL2 is indeed the smallest and most consistent cluster in which genetic analysis more often yielded negative results, and ECG parameters more frequently indicated increased QRS duration, LBBB, and markers of delayed and slower ventricular depolarization. This second cluster showed very few arrhythmic events during an eight-year follow-up and was, in fact, devoid of P/LP variants of well-known arrhythmic genes such as DSP, PKP2, and LMNA. This held true even if more than half of the patients in CL2 had LVEF $< 35\%$. It is of interest that the main ECG clustering variables identified, all pertained to the same sphere of indicators of Left Ventricle (LV) remodeling (i.e., hypertrophy/dilatation): patients in CL2 significantly more often met LVH Cornell criteria, had more delayed intrinsicoid deflection, more

often LBBB and wider QRS. It could be hypothesized that a larger LV mass (hypertrophy criteria, increased time for the electric signal to go from the endocardium to the epicardium) and fewer indicators of myocardial fibrosis (high voltages) in the field of DCM might protect from major arrhythmias through a less compromised myocardial muscle. On one hand, our results are in line with the recently published "Madrid Genotype Score" which proved to be an accurate tool to predict genetic test positivity in DCM [207]. In fact, the absence of LBBB and low voltages are predictors of a positive genetic test in this score. Complementarily, we found that high voltages and LBBB identify a cluster of gene-elusive DCMs. On the other hand, in a previous analysis of our group focused on ECG in DCM population [215] the only presence of true LBBB did not show a prognostic role in terms of HF and arrhythmic events. However, true LBBB is not the only parameter representing cluster 2, and other CL 2-parameters (LVH by Cornell criteria, S nadir, high QRS voltages) were in the same manuscript described as protective for MVA/SCD and LVH also for overall survival. One could wonder whether CL2 patients are truly idiopathic DCMs. Although the diagnosis was properly based on the exclusion of confounding environmental factors, it cannot be excluded that CL2 patients represent a cohort of patients in which the role of these factors is not negligible.

Clinical implications

Despite a strong prognostic role herein identified, the literature is poor of studies that investigate the role of ECG parameters in DCM and rather rich in second- and third-line tests in search for deep cardiomyopathy phenotyping. It would be worth reconsidering the wealth of information ECG can give on DCM patients and further exploring its predictive role in future studies, especially given we all already use this information in clinical practice. While machine learning approaches are fascinating, it may be often difficult to integrate them into clinical practice firstly because they need a wide range of variables and secondly because their result could be difficult to predict based on clinical experience alone. With this in mind, we believe the few variables we identified, and their immediate collection, have the merit of being easily applied by all clinicians, without the need for complex calculations. One of the implications of such an easily applied model is that also cardiologists not working in cardiomyopathy referral centers can use it to estimate the priority with which they need to refer a suspected DCM patient to a tertiary cardiomyopathy clinic. DCM is a widely heterogeneous condition, that often challenges the clinician especially upon first patient evaluation when it may be difficult to foresee which path will the disease take. In the future, the findings of this study could be put together with known prognos-

tic models to better stratify patients with DCM from their very first medical contact.

Limitations

Although sufficient for statistical analysis, the number of patients included in this study is limited, and further testing of the model in larger cohorts should be performed. Furthermore, although validated in an external cohort, the bias of enrolling patients from cardiomyopathy referral centers cannot be excluded. Furthermore, enrolling only genotyped patients from DCM referral centers introduced a selection bias. The limited patient number might also explain why while CL2 was homogeneous in phenotype and genotype, CL1 did not allow a correlation between ECG data and underlying genotype. A larger population would make possible a deeper characterization of this cluster. Nonetheless, the identification of a low-risk cluster is an important steppingstone for the clinical cardiologist.

3.2.5 *Conclusion*

Unsupervised clustering analyses using data collected at the first evaluation of DCM patients robustly identified a low-risk cohort of patients with distinctive ECG characteristics, possibly expression of a different underlying pathological process. The implementation of this model with simple ECG parameters in clinical practice will help in identifying from the start patients at lower risk for arrhythmic outcomes.

3.3 TREATMENT EFFECT PHENOTYPING

3.3.1 *Introduction*

This work focuses on advancing personalized treatments by addressing the limitations of the traditional one-size-fits-all approach, specifically in time-to-event outcomes. While existing methods often overlook how treatment effects change over time, this study proposes a method to identify dynamic treatment effect patterns, allowing for more personalized treatment decisions.

The proposed method estimates the Conditional Average Treatment Effect (CATE) over time, taking into account potential time-dependent effects. This allows for the identification of distinct treatment response patterns, or phenotypes, which can then be linked to individual patient profiles. The rationale is that by providing information on treatment effects across the entire time horizon, medical experts can make more informed, personalized decisions, such as determining whether a patient has an early or late response to a treatment or if the effect of a treatment increases or fades over time. The method combines survival neural networks and FDA techniques, using a Logistic-Hazard neural network (Nnet-survival) to model the hazard function non-parametrically. To achieve smooth estimates of CATE, the neural network's predictions are interpolated with natural cubic splines, similar to how smoothing is achieved in FDA. This allows for a flexible and robust modeling of time-dependent effects and interactions between covariates. Functional clustering is then applied to group and classify the time-varying CATE estimates, identifying relevant treatment effect phenotypes.

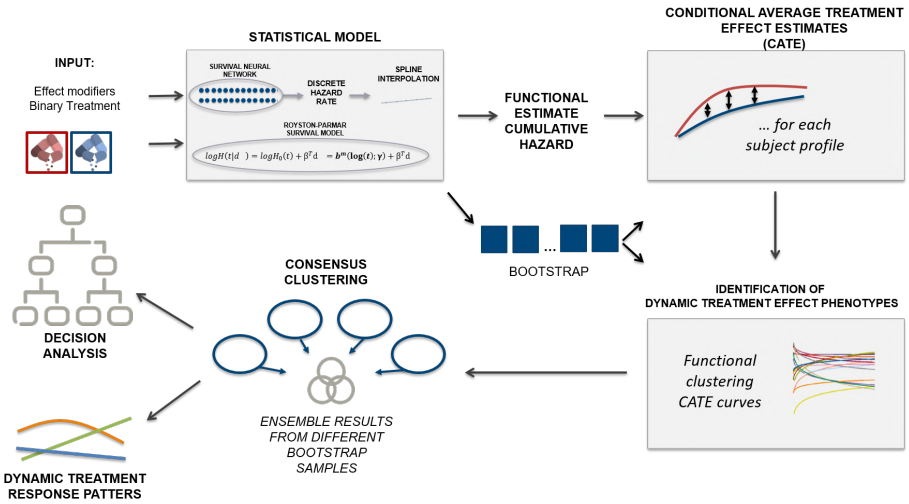


Figure 19: Summary of the proposed method to obtain dynamic treatment effect phenotypes.

3.3.2 *Methods*

Notation and Formal Framework

We let D and C be the event of interest and censoring time respectively. For each subject, we only observe the couple $\min(D, C)$, $\delta = \mathbf{1}(D \leq C)$. The binary random variable indicating the treatment/exposure is denoted by Z. Let $\mathbf{X}$ be a p-dimensional vector of observable covariates with $p > 1$. We define as $D^{(1)}$ the potential outcome if the treatment/exposure is received and $D^{(0)}$ the corresponding potential outcome without treatment/exposure. We are interested in estimating the CATE at each time $t \in [0, w]$ on a subset of the covariate vector $\mathbf{X}$, denoted as $\mathbf{X}_1$, that contains the treatment-effect modifiers :

$$\tau_{\mathbf{x}_1}(t)^{[0,w]} = E\{f[g(D^{(1)}; t), g(D^{(0)}; t)]|\mathbf{X}_1 = \mathbf{x}_1\} \tag{10}$$

where $g(\cdot)$ can be either the survival, hazard, or cumulative hazard function and $f(\cdot)$ is a measure of effect, typically either the ratio or the difference. In the following, simplifying the notation, we omit the reference to the specific time horizon of interest, denoted as $[0, w]$, in $\tau_{\mathbf{x}_1^{[0,w]}}(t)$.

To identify the CATE, we need to assume the treatment groups are *conditionally exchangeable*. If this is not verified, a propensity score method such as matching or Inverse Probability of Treatment Weighting needs to be applied to ensure conditional exchangeability.

The steps of the method to identify treatment-effect phenotypes consist of 1) estimating the CATE as in eq. (10), 2) identifying the grouping structure of the CATE curves through an unsupervised clustering method 3) characterizing the different behaviors in terms of response to the treatment, and 4) mapping the different subject profiles to their corresponding treatment-response behavior.

Estimation of functional CATEs

In the Royston-Parmar Survival Model (R-PSM) [216], the logarithm of the baseline cumulative hazard function is modeled as a natural cubic spline function of log time. The model allows for flexibility in the form of the baseline hazard by adjusting the number of internal knots and accommodates heterogeneity in treatment effects by including covariate-treatment interactions. It can also model time-varying effects by modeling the spline coefficients in the function of the covariates for which a time-varying effect is desired. The parameters in the model can be estimated using maximum likelihood and their uncertainty can be evaluated using standard asymptotic theory.

A model-free approach consists of using a survival neural network that estimates the hazard function. Specifically, we consider the Nnet-Survival method, also called Logistic-Hazard, which parametrizes the discrete-time hazard rate with a neural network and optimizes

a survival likelihood expressed in terms of the discrete hazards non-parametrically [187]. In the Nnet-Survival method, follow-up time is divided into h intervals which are left-closed and right-open. The contribution of a generic time interval j to the overall log-likelihood is:

$$\sum_{i=1}^{d_j} = \log\left(h_j^i\right) + \sum_{i=d_j+1}^{r_j} \log\left(1 - h_j^i\right) \tag{11}$$

where r_j is the number of subjects at risk before the beginning of the interval, d_j is the number of subjects experiencing the event during the interval and, h_j^i is the hazard for an individual i during the time-interval j. The loss function in eq. (11) comes from classic discrete-time survival models, and its use in a neural network context is well justified by survival analysis theory. Furthermore, it naturally incorporates a time-varying baseline hazard rate and time-varying effect, since each time interval output node is fully connected to the last hidden layer's neurons. As architecture, we used a fully connected network with 2 hidden layers of 32 units each, and ReLU as a non-linear activation function. The neural network gives, for each subject profile, a h-dimensional output corresponding to a discrete set of hazard rates, one for each interval. We then propose to interpolate the hazard curves using natural cubic splines. In this way, similar to the previous model we obtain a smooth estimate of $H(t|d)$ that depends in this case on a non-linear function of the covariates x_1 and the treatment indicator z. This neural network was implemented in `python` using `pycox` [185] and `pyTorch` [132].

From both the R-PSM and the Spline Nnet-Survival (SNnet-S), we obtain a functional estimate of the CATE as defined in eq. (10), by predicting the measure of effect of interest, $g(\cdot)$, for each combination of the covariates x_1 and by comparing it between the two treatment strategies $z = 0$ and $z = 1$ using the chosen function $f(\cdot)$. The estimation of the CATEs is only the first step of the procedure that leads to the identification of the treatment-effect phenotypes. To address uncertainty and enhance the overall reliability of our findings, we employ resampling techniques to quantify the variability in CATE estimates. For the R-PSM method, a simulation approach based on parametric bootstrap [217] is applied, while for the neural network, a non-parametric bootstrap is used. This means that for every combination of the covariates $\mathbf{x}_1$ under investigation, we generate a distribution comprising B estimates of CATEs, that we will employ in the following step for the identification of the treatment effect phenotypes.

Identification of treatment effect phenotypes over time

Functional clustering is used to find a grouping structure for $\hat{\tau}_{\mathbf{x}_1}^{(b)}(t)^{[0,w]}$, $b = 1, \ldots, B$ for each combination of the covariates $\mathbf{x}_1$ of interest in

order to aggregate profiles of individuals in clusters with a common response to the treatment exposure over time. The functional clustering procedure is repeated on each bootstrap sample to take into account the uncertainty in the estimation of the CATEs and improve the stability of the procedure. For FDA clustering, we used the implementation in the `R` package `fdacluster` [218]. Specifically, we used functional K-means clustering, choosing the L2 distance between the CATE estimates of the different subject profiles:

$$d_2\{\hat{\tau}_l^{(b)}(w), \hat{\tau}_m^{(b)}(w)\} = \sqrt{\int_0^w \{\hat{\tau}_l^{(b)}(s) - \hat{\tau}_m^{(b)}(s)\}^2 ds} \tag{12}$$

for $b = 1, \cdots, B$ and each couple of subject profiles l, m.

To aggregate clustering results from bootstrap samples, we used two approaches: majority voting and consensus clustering, which aims to maximize the similarity between clusters. Medoid consensus clustering [219] selects the final clusters from the set of base clusterings, while soft consensus clustering [220] allows objects to belong to multiple groups with varying degrees of membership. K-means clustering requires choosing the number of clusters. The silhouette values are typically employed for this task. Alternatively, the re-sampling procedure through bootstrap allows us to consider the mean internal agreement between the results of the clustering on the different bootstrap samples to assess the stability of the results with different numbers of clusters. Finally, clusters are interpreted through the visual inspection of the clustering consensus centroids. These were obtained as the functional median obtained using the Modified Band Depth [221, 222] of the CATEs contained in each final consensus cluster.

3.3.3 *Simulation Study*

Design

We performed a simulation study to evaluate the performance of the proposed method. In the first part of the simulation, we compare the performance of the SNnet-S with the R-PSM. Secondly, we wanted to assess the stability of the identification of treatment effect patterns through functional clustering. We considered a vector of 5 binary effect modifiers, and we specified a data-generating model based on the spline-based parametric survival model with time-varying effects and several interactions among the covariates. According to this model, the baseline hazard was specified considering a spline with 1 knot placed at $t = 10$. 100 datasets were simulated under four different sample size scenarios (n=1500, 5000, 10 000, 50 000). As a measure of effect in the simulation, we considered the ratio between cumulative hazard.

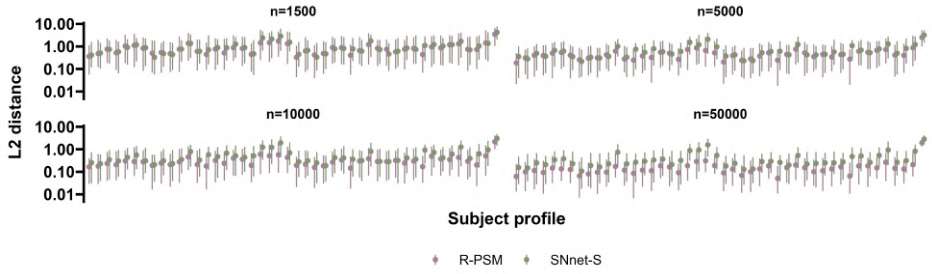


Figure 20: Comparison of the median L2 distance with 95% CI between the estimated cumulative hazard of a correctly specified R-PSM (pink) and SNnet-S (green), relative to the true data generating model.

Results

The L2 distance between the cumulative hazard of the true data-generating model and the one estimated with either the correctly specified R-PSM or the SNnet-S approach was calculated for each combination of the covariates and binary treatment indicator. The results for the six scenarios are reported in Figure 20. For all sample sizes and all subject profiles, the SNnet-S and the R-PSM reach a similar performance. It is important to note that the performance of R-PSM is conditioned on the fact that we are able to select the correct model.

Once established the satisfying performance of the SNnet-S approach, we continued with this approach and performed the functional clustering considering different numbers of clusters $nclust = 2, 3, 4$. 100 was used as the number of bootstrap samples. As previously reported, to ensemble the clustering results obtained in the different bootstrap samples a naive majority vote and consensus clustering methods were considered. Specifically, here we considered three medoid consensus methods and two soft least square consensus methods that differ for the (di-)similarity measure used: Euclidean, Manhattan, and Rand are considered for the medoid consensus, and Euclidean and Manhattan are considered for the least square soft consensus.

The results of the clustering obtained on cumulative hazard ratio curves of the data-generating model are considered the "gold standard" and they are compared in terms of agreement across the different ensemble bootstrap methods employing the Jaccard Index and the Corrected Rand (CRand) Index. In Figure 21 we can observe that the agreement increases with the sample size. Importantly, the results are robust to the choice of the clustering consensus method. The agreement is higher for the two clusters, and this is due to the specific data-generating model used. In particular, considering two clusters, the method achieves an agreement with the "true" clustering above 0.90 according to both indices and all clustering consensus methods on the scenario $n = 50\,000$. In Table 10 we report the mean agree-

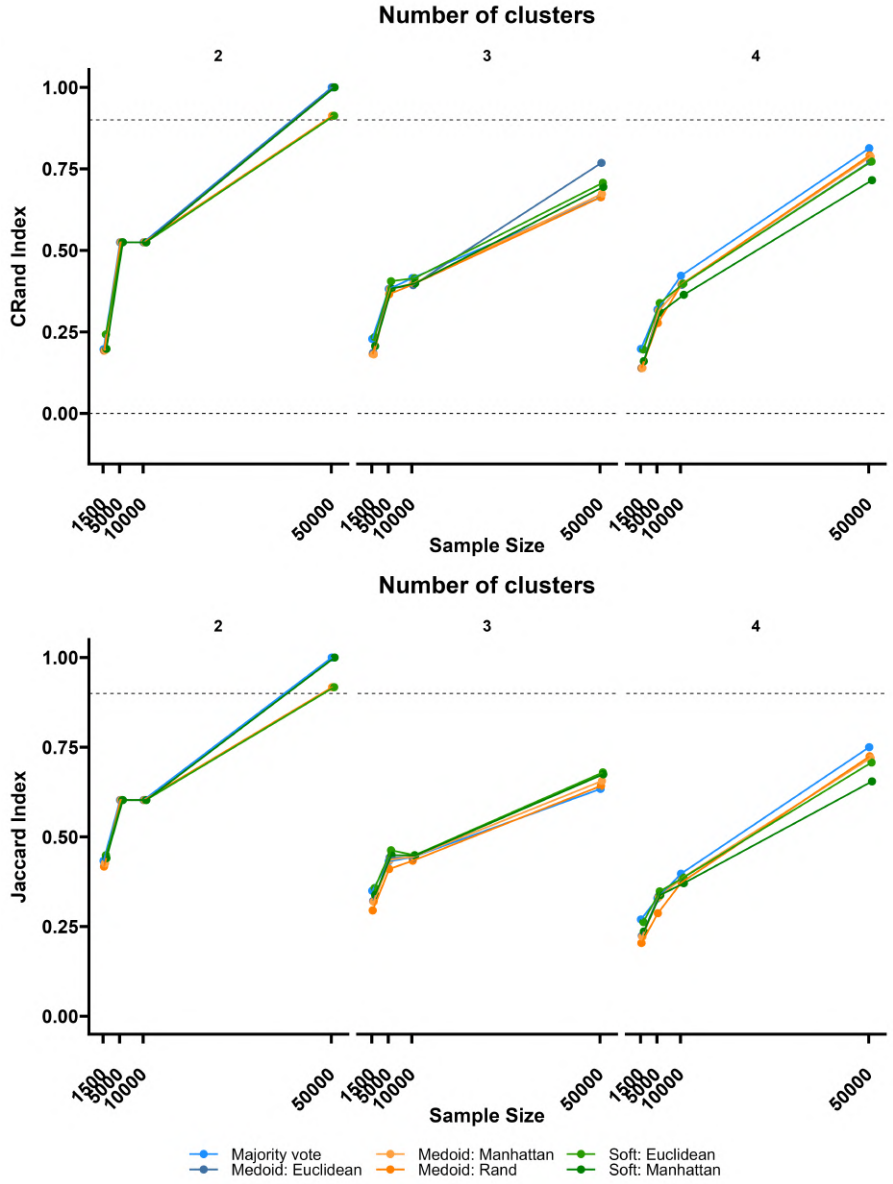


Figure 21: Comparison of the agreement according to the CRand Index (top panel) and the Jaccard Index (bottom panel) according to different bootstrap ensemble methods.

Table 10: Concordance of clustering results between bootstrap samples.

Sample size	Number of clusters	CRand Index	Jaccard
1500	2	0.06	0.38
	3	0.05	0.25
	4	0.04	0.19
5000	2	0.22	0.45
	3	0.19	0.32
	4	0.16	0.25
10000	2	0.30	0.49
	3	0.25	0.36
	4	0.21	0.28
50000	2	0.57	0.65
	3	0.46	0.50
	4	0.41	0.41

ment among the clustering results in the 100 bootstrap samples. As expected, a higher internal agreement among clustering samples corresponds to a higher final performance of the method. For two clusters and $n = 50\,000$, the mean internal agreement is above 0.5.

3.3.4 *Discussion and Conclusions*

This research introduces a novel method to characterize different treatment responses in a survival setting, focusing on effect heterogeneity and time-varying effects. Most existing methods categorize treatment outcomes as "beneficial," "null," or "harmful," but there is a broader range of possibilities. It is crucial to understand the dynamic response to treatment and summarize to some extent which are the most common treatment-effect phenotypes over time and to which subject profiles they can be attributed.

Our method employs a two-stage procedure: first, we estimate the conditional average treatment effect over time using a DL model that accommodates interactions and non-proportional hazards. We demonstrated the effectiveness of survival neural networks and subject-specific hazard curve smoothing through simulation studies. The second stage involves clustering the curves representing conditional treatment effects to summarize the most common treatment-effect phenotypes over time and their associated subject profiles.

When interpreting the results, two aspects are important: first, the treatment effect phenotypes can be understood by visualizing the medoids of the clusters. Second, the mapping between the effect mod-

ifiers space and the phenotypes can be represented by a decision tree, which can be easily used by domain experts to subjects' assignment to a specific phenotype. However, the interpretation of the effect of specific covariates on the CATEs is outside the scope of this study.

A limitation is represented by the need to choose the number of treatment-effect phenotypes. Even though there are established methods coming from the clustering literature, in this specific context it is also important to confirm the interpretability of the results from a domain-specific point of view. Moreover, the method relies on the choice of a specific clustering algorithm. Although we have considered functional K-means, it would be possible to use any functional clustering algorithm. Indeed, consensus clustering is independent of the specific clustering method used, and it could in theory also be used to combine the results obtained from different algorithms without the need to choose one algorithm over the other.

4

ANOMALOUS AORTIC ORIGIN OF THE CORONARY ARTERY DETECTION

This section of the thesis is based on work conducted during a visiting period at the Bern University Hospital (Switzerland), under the supervision of Dr. Isaac Shiri in the research group *Artificial Intelligence in Cardiovascular Medicine Laboratory*. The study presented here has not been published yet but is currently under review.

4.1 INTRODUCTION

Anomalous Aortic Origin of Coronary Arteries (AAOCA), introduced in Section 1.2.4, is a rare congenital heart condition that presents in various forms [223]. Most AAOCA cases, particularly those involving a pre-pulmonic or retro-aortic course of the anomalous vessel, are generally considered low-risk [224, 225]. However, AAOCA with an intramural or interarterial course (i.e., where the vessel runs between the aorta and pulmonary artery) is classified as "malignant" due to the risk of ischemia and adverse cardiac events [225].

Accurate detection is crucial for proper management of the condition [225–227]. AAOCA can be discovered incidentally during imaging studies for other conditions, such as CAD [228]. As coronary CCTA has become a primary non-invasive imaging technique for CAD, the number of detected AAOCAs is increasing [76, 224, 226, 227]. However, the condition may still be missed, especially in smaller centers where there is less experience with this rare anomaly. Additionally, imaging fellows may lack the familiarity needed to identify and correctly classify AAOCA as high-risk, leading to uncertainty in diagnosis. Given these challenges, there is an unmet need for automated tools to analyze CCTA images to reduce the risk of overlooking high-risk AAOCA cases. Such tools could also support large-scale analyses to improve the detection in retrospective or prospective datasets and enhance our understanding of the relationship between AAOCA variants and clinical outcomes.

AI-based tools have been developed for 3D-CCTA image analysis, including automated coronary centerline detection, coronary segmentation, automated classification and measurement of plaque, and automated reporting, but none of these tools are available for automated AAOCA detection [229, 230]. Here we developed a fully automated AI-based screening tool for detecting and classifying AAOCA in 3D-CCTA images.

4.2 METHODS

The design and reporting of this study follow different dedicated guidelines for AI applications in medical imaging, including CLAIM (Checklist for Artificial Intelligence in Medical Imaging) [231], STARD-AI (Standards for Reporting of Diagnostic Accuracy Study-AI) [232], and MINIMAR (Minimum Information for Medical AI Reporting) [233]. Items of the above-listed guidelines documents have been jointly considered appropriate for the development, validation and testing of the AI-based model for automated AAOCA detection and classification in 3D-CCTA. Figure 22 shows an overview of the entire study implementation, including the datasets, examples of CCTA images, the different networks, and a summary of the results.

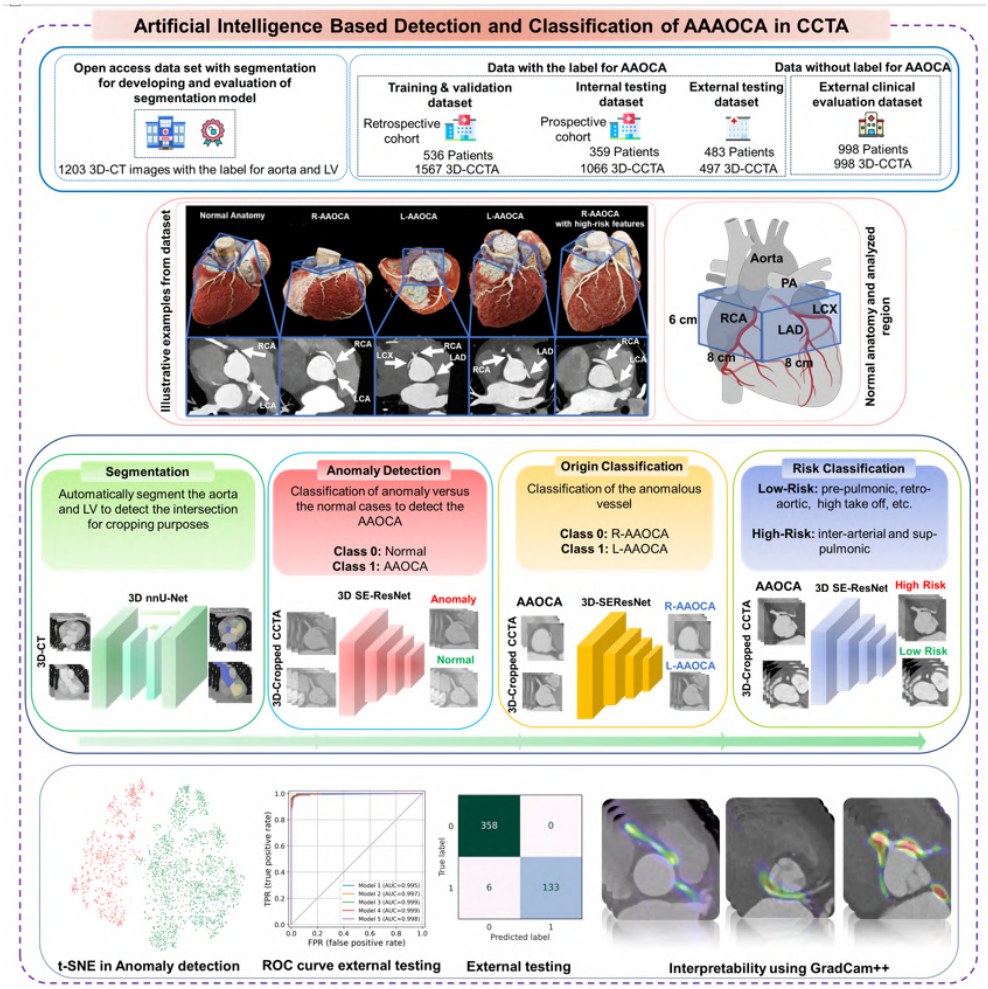


Figure 22: The flowchart of the current study shows an overview of the entire study implementation, including the datasets, examples of CCTA images, the different networks, and a summary of the results. RCA: Right coronary artery, LAD: left anterior descending, LCX: left circumflex, PA: Pulmonary Artery

4.2.1 *Datasets*

This study utilized multiple datasets for different tasks. Initially, to develop the segmentation model, we employed 1203 open-access CT images that included ground-truth segmentations of the aorta and LV [234]. Additionally, data from three different centers were used to develop and evaluate the detection and classification of AAOCA. Bern University Hospital provided both a retrospective cohort (training dataset) and a prospective cohort (internal testing dataset), with data collected before and after April 2020, respectively. Zurich University Hospital contributed a single dataset used as an external testing dataset. From these two centers, the datasets contained both normal and AAOCA cases, with labels specifying the type of anomaly, the anomalous coronary artery, and its anatomical course, as determined by two cardiologists specialized in CCTA. The third dataset was an open-access 3D-CCTA dataset from Guangdong Provincial People's Hospital, acquired from April 2012 to December 2018 [235]. This dataset was unlabelled and was used to assess the model's screening functionality in real-world scenarios (external clinical evaluation dataset). Detailed information for each dataset is provided in Table 16 in the supplementary material. Information regarding the open-access dataset and segmentation datasets is available online [234, 235]. All procedures in studies involving human participants adhered to the ethical standards of the institutional and/or national research committee, the 1964 Helsinki Declaration, and its subsequent amendments or comparable ethical standards. The Bern cantonal ethics committee approved the study design (KEK 2020-00841, Registry for Invasive and Non-invasive Anatomical Assessment and Outcome of Coronary Artery Anomalies (NARCO) ClinicalTrials.gov: NCT04475289 and Clinical Utility and Outcome Prediction of Cardiovascular Computed Tomography (PREDICT-CT) KEK 2021-0058 NCT04827316). Participants in the study provided written informed consent prior to any data collection, and all imaging data were anonymized.

4.2.2 *Segmentation Model*

In AAOCA, the origin and proximal course of the anomaly are critical. Therefore, we decided to focus our analysis exclusively on this region compared to the entire 3D-CCTA cardiac image. We developed a fully automated segmentation model to automate cropping in this area. Given the high variability in the anatomy of the aorta and to ensure robust cropping that does not fail in different datasets, we segmented both the aorta and the LV. We employed the nnU-Net [236] network with deep supervision and extensive data augmentation to develop this segmentation model. The models were trained using a 5-fold approach, and the ensemble of 5 models was used for inference on the

test sets. This ensemble model was then applied across the rest of the dataset for which we had no labeled segmentation (dataset for classification from training, validation, internal and external testing, and external clinical evaluation dataset).

After segmenting the aorta and LV, the region growing between the two segmentations was performed, and the center of the intersection between the two segments was identified. The point was then shifted 1 cm to the right and upward (to have more aorta in the crop images due to its curve anatomy), and a box with dimensions of 8×8×6 cm^3 was fitted over the image, and the region of interest was cropped. This cropping approach ensures that all regions relevant to the origin and course of the artery are included.

4.2.3 *Classification Model*

Image preprocessing

Before being input into the classification models, the cropped images underwent different steps of preprocessing. These steps were defined and executed solely on the training and (internal)-validation set, and then, the same procedures were applied to the testing dataset to maintain consistency and prevent any data leakage. Initially, cropped images were resampled using spline interpolation to uniform dimensions of 215×215×85, determined by the median values in each direction from the training and (internal)-validation dataset. To minimize noise and remove outlier intensities, the resampled images were clipped to a Hounsfield Units range of -1024 to 1024. Lastly, to reduce the dynamic range further, the image intensities were discretized into 256 levels (ranging from 0 to 256) and subsequently normalized to a scale from 0 to 1 with a min-max scaling approach.

Data augmentation

We implemented various data augmentation techniques, including random noise, blurring, and contrast, to mimic different presentations in 3D-CCTA images from different scanners and centers. Moreover, to increase the robustness of our model under conditions of motion and step-and-shoot acquisition in older CCTA scanner generation, which can significantly change image appearance, we simulated these conditions and included them in our augmentations. Examples of different image augmentations are provided in Figure 38 in the supplementary material.

Deep-learning modeling

A 3D Squeeze-and-Excitation residual Network (SE-resNet) consisting of 154 layers [237] was trained for various detection and classification

tasks. To train the networks, we split the training dataset into training and (internal)-validation sets with a ratio of 90:10% (Figure 23). In the data-splitting process, we made sure that the images of the same patient were present only in training or (internal)-validation. The learning rate was initially set at 0.001 and adjusted based on a cosine annealing scheduler [131], ensuring steady convergence. The network was trained over 300 epochs, adopting binary cross entropy as the loss function, and early stopping was implemented to prevent overfitting. The best model was selected based on its performance in reducing training and (internal)-validation loss and its overall accuracy on (internal)-validation datasets. We used different metrics such as accuracy, sensitivity, and specificity, evaluating the models comprehensively during the (internal)-validation phase, ensuring that the chosen model minimized loss and maximized predictive reliability.

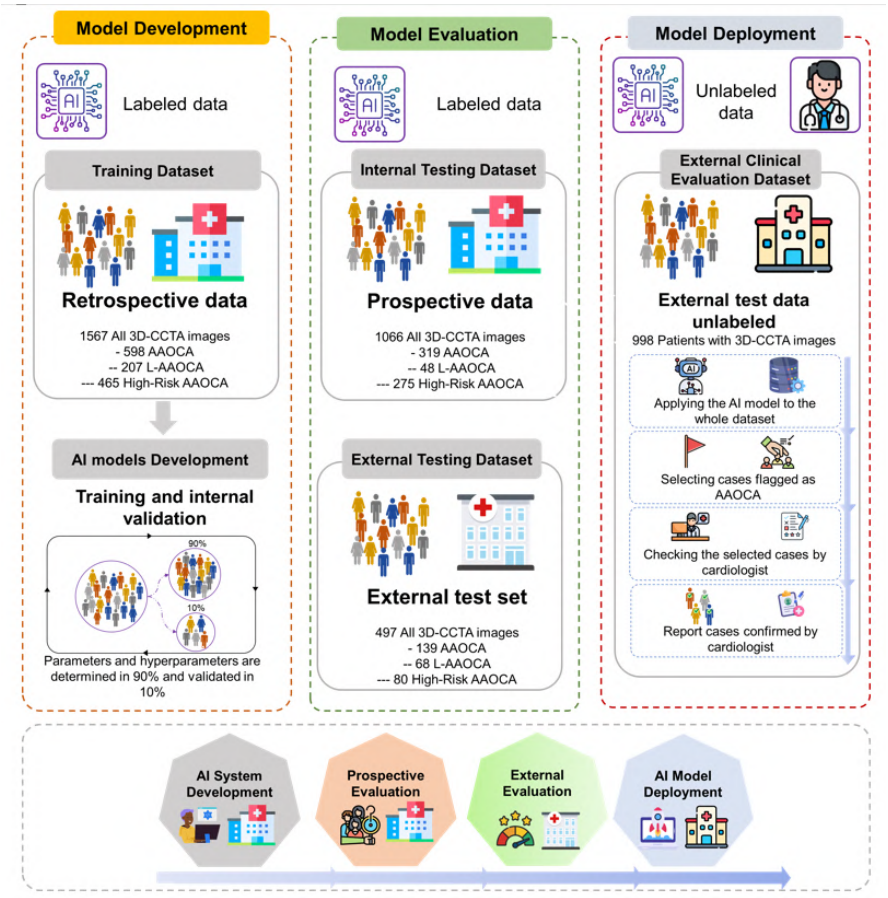


Figure 23: Summary of model development and evaluation throughout the entire study: the model was developed using a retrospective dataset from the training and internal validation dataset and tested on a prospective cohort of the internal testing dataset from the same center. Models were then externally validated (external testing dataset). We finally simulated a real-world scenario using an unlabeled dataset (external clinical evaluation dataset).

Different classification tasks

In our study, we tackled three distinct tasks:

- **Anomaly detection**: distinguishing between normal cases and those with AAOCA.
- **Origin classification**: classifying the anomalous vessel into either right (R-AAOCA) or left (L-AAOCA), including anomalies of just the left anterior descending or the left circumflex coronary artery.
- **Risk classification**: classifying the AAOCA risk as either low-risk anatomy (characteristics like pre-pulmonic, retro-aortic, and high take-off) or high-risk anatomy (features such as inter-arterial and sub-pulmonic courses).

The training of all models was conducted on the training dataset, where all model parameters and hyperparameters were established and used without any changes in the other datasets. For testing, we employed 3 different datasets (Figure 22 and 23) to ensure robustness and generalizability. The internal testing dataset served as an internal testing set, whereas external testing and clinical evaluation datasets provided real-world external test scenarios to assess model performance in different settings.

For the coronary anomaly detection, the model was trained from scratch with a Kaiming weight initializer. For the other two classifications (origin and risk classification) models, which had smaller training sets and included only the AAOCA data subset, we used the pre-trained network from the anomaly detection model.

Evaluation of the models

Performance evaluations for all models were carried out using a variety of metrics in the internal and external testing datasets. We reported different classification metrics, such as AUC, sensitivity, and specificity for each task. Due to the absence of labeled AAOCA cases in the external clinical evaluation dataset, we evaluated the model's screening performance for anomaly detection, origin, and risk classification in real-world scenarios. We applied the same preprocessing steps and used the trained model on the entire dataset. We then selected cases flagged as AAOCA by the AI models. These cases were subsequently reviewed by two cardiologists experienced in CCTA. A final report was compiled based on these expert assessments to provide a detailed evaluation of the model's performance in detecting and classifying AAOCA in a real-world setting without prior labeling. Figure 23 presents a summary of model development and evaluation throughout the entire study. Supplemental Figure 39 illustrates different strategies for model development and evaluation, which were implemented within our dataset. More specific:

- Strategy 1: Model development was performed on the training dataset; the models were evaluated on the internal and external testing dataset with labeled cases. The external clinical testing dataset was used to evaluate the true and false positives, as the labeling was not available for this dataset.
- Strategy 2: Model training was performed on the entire dataset from Bern University Hospital. The labeled dataset from Zurich University Hospital served as an external testing dataset. The unlabeled open-access CCTA dataset (Guangdong Provincial People's Hospital) was used for external clinical evaluation, similar to Strategy 1.
- Strategy 3: Model training was performed on the entire datasets with labels, including data from Bern and Zurich University Hospitals. Following the previous strategy, external model performance was evaluated in the unlabeled dataset (external clinical evaluation dataset).

We report the results of Strategy 1 (Figure 23) in the main manuscript, while the results of Strategies 2 and 3 are reported in the Supplementary material. These additional strategies were explored to enhance the performance of the final model using the entire labeled dataset in different approaches. Supplemental Figure 40 shows the different options for using the developed model in real clinical settings, from fully automated to semi-automated (physician in the loop) approaches.

Ensembling, interpretability, feature visualization

Given the numerous possibilities for network architectures and configurations that could be explored, we selected our model based on initial experiments conducted solely on the training and internal validation sets. To ensure the reproducibility of our results and avoid false discovery and considering the limited computational power, we reported the model performances across five different trainings named 5-fold models (the 5 different models were trained using the same training– (internal)-validation split). Then, we obtained an ensemble model whose predictions are the average of the single models' predictions. Moreover, to make the models more interpretable and to understand the features influencing their decisions, we implemented the Grad-CAM++ [238] (Gradient-weighted class activation mapping) algorithm. This allowed us to visually highlight which areas of the images were most significant in determining the outcomes. Additionally, we used t-distributed Stochastic Neighbor Embedding (t-SNE) to visually represent the dataset based on the embedding features learned by the model and to see how these features discriminate between different data classes.

Model development and data availability

All developed code and models are made publicly available on our AI-CVI laboratory's GitHub page (`https://github.com/AI-in-Cardiovascular-Medicin` `AAOCA`). All data preprocessing and model development were conducted using Python and different libraries such as `ITK`, `PyTorch`, `TorchIO`, and `MONAI` (more details are provided on GitHub). All computational was performed on high-performance servers equipped with 3 A100 GPUs, 250 CPU cores, and 1 TB of VRAM. The dataset used for segmentation model development and clinical evaluation is publicly available [234]. The datasets from Bern and Zurich University Hospitals (training, internal validation, internal and external testing set) are not shareable due to a lack of dedicated ethical approval for this purpose. The dataset from Guangdong Provincial People's Hospital (external clinical evaluation dataset), which can be used to test and evaluate different segmentation and classification models, is publicly available in [235].

4.3 RESULTS

4.3.1 *Dataset*

Figure 23 shows descriptive information about the individuals and images in each dataset, and Supplemental Table 16 provides more details of the different datasets used for model training and evaluation. For the AAOCA detection, 2376 patients (4128 CCTA images) were included, with 335 AAOCA patients (1056 CCTA images). Out of the entire dataset, 998 patients (998 CCTA images) did not have labeled cases for AAOCA, which corresponds to the external clinical evaluation dataset. For the anomalous coronary artery classification task, 328 patients (1029 CCTA images) were included, with 133 L-AAOCA patients (323 CCTA images). For the risk classification tasks, 327 patients (1026 CCTA images) were included, with 226 high-risk anatomy patients (820 CCTA images).

4.3.2 *Segmentation*

The segmentation model for the aorta and LV achieved a mean dice score of 0.89 (0.83 for myocardium, 0.90 for LV cavity, and 0.93 for aorta) in 5-fold cross-validation. Using this segmentation model, we cropped all images to the desired size. Although we needed only a rough segmentation for cropping purposes, we checked all segmented and cropped images across different datasets and found no mis-cropping. Examples of segmentation were provided in the Supplemental Figure 41.

Table 11: Summary of different classification metrics for the ensemble models for different test datasets. Anomaly detection: distinguishing between normal cases and those with AAOCA; Origin classification: classifying the anomalous vessel into either the right (R-AAOCA) or left (L-AAOCA); Risk classification: scoring the AAOCA risk, classifying it as either low-risk or high-risk anatomy.

	Anomaly detection		**Origin classification**		**Risk classification**	
	Internal testing	External testing	Internal testing	External testing	Internal testing	External testing
AUC	0.998	0.999	0.999	0.999	0.999	0.996
Sens.	0.988	0.957	0.938	0.956	0.989	0.963
Spec.	0.989	1.000	1.000	1.000	1.000	0.964
F1-score	0.981	0.978	0.968	0.977	0.995	0.969
PPV	0.975	1.000	1.000	1.000	1.000	0.975
AUPR	0.967	0.969	0.947	0.978	0.999	0.960
Acc.	0.989	0.988	0.990	0.978	0.990	0.963

4.3.3 *Classification*

Summary results of different detection and classification tasks for two different testing sets (internal and external testing datasets) for the ensemble model of 5-fold are presented in Table 11. The AUC of the different models was more than 0.99 across all testing datasets and models. For the detection of anomalies, a sensitivity of 0.99 and 0.95 was achieved for the internal and external testing datasets, respectively. The specificity for anomaly detection and origin classification was higher than 0.99 across all testing datasets. Sensitivity and specificity in risk classification were both higher than 0.96 in both testing datasets. More detailed results for all tasks and test datasets, including fold-specific metrics and average performance, are provided in Supplemental Tables 17-19.

The Receiver Operating Characteristic (ROC) curves for the five different folds and the ensemble model for different detection and classification tasks across different testing datasets (internal and external testing datasets) are provided in Figure 24. Supplemental Figure 42 shows the ROC curve for the mean of the five folds. Figure 25 presents the confusion matrix for the classification models in different detection and classification tasks across different testing datasets for the ensemble model, using a cut-off of 0.5 based on the training and internal validation dataset. Supplemental Figures 43-45 illustrate the confusion matrices for classification in different detection and classification tasks with cut-offs ranging from 0.1 to 0.5. These figures

demonstrate that by decreasing the cut-off, the number of true and false positives could increase to some extent, allowing for a higher sensitivity and lower specificity model depending on clinical use.

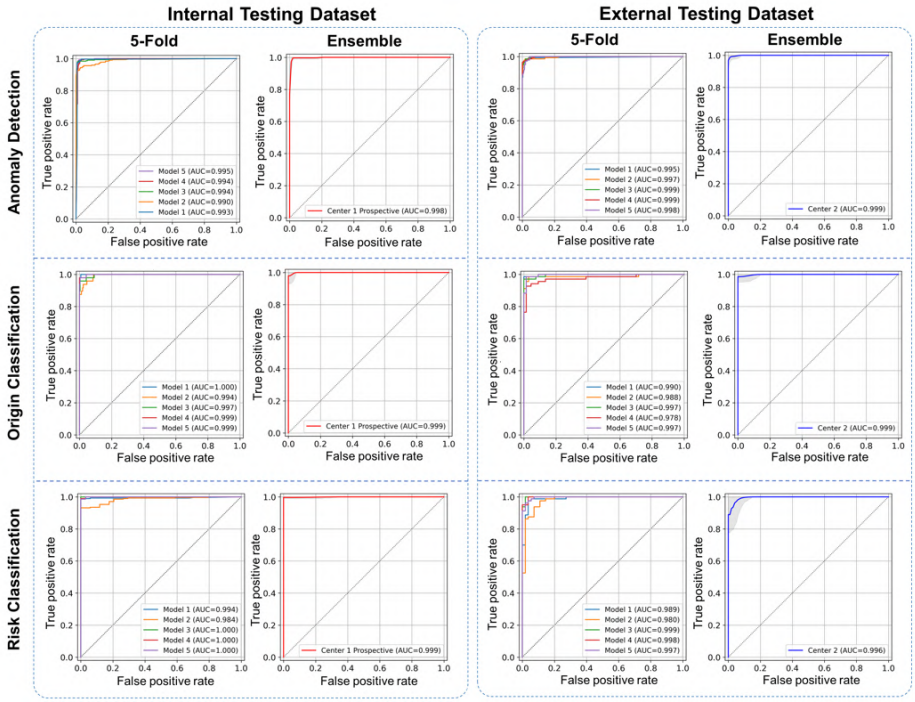


Figure 24: ROC curves including the 5 different folds and the ensemble of 5 models across different tasks for various test datasets. Confidence intervals for the ensemble models were computed with the bootstrap method (10 000 iterations).

4.3.4 *Interpretability*

Figure 26 illustrates different cases, including normal and AAOCA, with various anomalies and the corresponding GradCam++ feature overlays on the CCTA images. As shown in this figure, the networks learned the exact pattern of the coronary arteries, and their activation maps indicate the location and course of different coronary arteries.

4.3.5 *Feature space visualization using t-SNE*

Figure 27 shows the 2D t-SNE maps of the latent features extracted from the last layer of one of the models trained for the anomaly detection task, with each point representing an image from different datasets. This figure was generated based solely on the feature set, without including any additional information such as labels, centers, or datasets in the plotting process. Labels and dataset types were used only for plotting and coloring purposes. As shown in Fig-

ure 27 - Panel a, there is no center or data-based clustering, indicating the model's generalizability without bias toward any specific center or database. Moreover, when colored based on normal and AAOCA cases, it shows two distinct clusters that differentiate between the two classes, with only a few cases misclassified. Figure 27 - Panels b-c display the latent features only for the AAOCA cases, this time colored by the type of anomaly (coronary artery origin and risk classification of their course). Looking at the color, it is possible to observe distinct clusters, suggesting that the network has automatically learned different patterns of **AOOCA!** (**AOOCA!**), even without explicit training for this task.

4.3.6 *Screening*

Figure 28 shows different cases from the external clinical evaluation dataset that were detected as AAOCA and confirmed by a physician. As shown in this figure, one high take-off (R-AAOCA, low-risk anatomy), two R-AAOCA (high-risk anatomy), and one L-AAOCA (left circumflex, low-risk anatomy) were detected and confirmed out of 24 cases flagged as anomalies out of 998 patients. Moreover, we highlighted challenging cases that were correctly identified as normal, including those highly affected by motion artifacts. Despite their high similarity to the different AAOCAs, the network could classify them correctly. Additionally, we present false positive cases, which were detected as AAOCA by the networks but were not confirmed as AAOCA by the physician. In all four true positive cases, origins and risks were correctly classified by the models.

4.3.7 *Real-world use case*

Supplemental Figures 46 and 47 show the results from Strategy 2, indicating a slight improvement in performance with no significant differences compared to Strategy 1. In Strategy 3, the entire labeled datasets were used for training. In both strategies, we did not find more positive cases from the external clinical evaluation dataset; however, the number of false positives decreased from 20 to 10 and 9 in the external clinical evaluation dataset in Strategy 2 and 3. The model with Strategy 3 was more likely to produce generalizable results in real-world scenarios across other centers, as it utilizes all labeled datasets.

4.4 DISCUSSION

In this study, we developed a fully automated tool to detect AAOCA, depict its origin and course, and classify high-risk anatomy. These tools could assist in real-time AAOCA detection of CCTA analysis and

help to reduce human error, and enable the prospective and retrospective analysis of large datasets, which could identify cases that might be missed during analysis. Accurate detection of AAOCA in CCTA is critical for optimal management, as those with high-risk anatomy are potentially leading to hemodynamic relevance [77, 78, 80, 239]. This is particularly important because CCTA is now established as a first-line noninvasive imaging technique for suspected CAD, leading to increased detection of absolute numbers in AAOCA, and therefore the differentiation of whether AAOCA is benign and a coincidental bystander or responsible for the patient's symptoms with anatomical high-risk is important to quickly detect for improving management. However, AAOCA may still be overlooked in routine clinical settings, especially as imagers may rarely comforted with this entity. This is emphasized by the fact that, although AAOCA is generally considered to have a very low prevalence, it may be more common than previously thought in the current area [76, 224, 226, 227, 240–243]. Additionally, the difficulty in correctly classifying anomalies as potentially malignant (interarterial course) versus benign needs expertise, which may be, lacking in smaller centers providing CCTA programs. Consequently, this AI tool could help to real time alert physicians about AAOCA and its potential risk. Furthermore, AAOCA has been associated with adverse events and sudden cardiac death, particularly during physical activity, as evidenced by autopsy reports [77, 78, 80, 239]. However, autopsy reports are associated with a selection bias towards potentially high-risk patients and do not represent the risk for living individuals with AAOCA [244]. Therefore, there is an unmet need to analyze large retrospective and prospective CCTA datasets, such as CONFIRM 2 [245], to assess the true risk of AAOCA when linked to outcomes. Our AI tool would facilitate the analysis of these large datasets, improving AAOCA detection and, by correlating findings with clinical outcomes, improving risk stratification based on AAOCA anatomy.

Although AI-based tools have been developed for various aspects of CCTA image analysis, none is currently available for automated AAOCA detection, especially due to the lack of proper ground truth. To develop a fully automated tool for the detection and classification of AAOCA, we implemented a two-step deep learning algorithm that includes segmentation/cropping and classification. Deep learning models were developed to detect anomalies and determine which coronary artery was affected, classifying the course as high or low-risk anatomy. The segmentation and cropping processes performed robustly across all cases without any failure. The classification model performance achieved high accuracy levels for detection and classifications in the internal and external test sets. Moreover, to simulate a real-world clinical scenario, we tested the model with an additional external dataset without any label for AAOCA, where the model

flagged a few cases as anomalies in a large dataset. These flagged cases were then reviewed and confirmed as anomalies by cardiologists, showing the model's effectiveness in identifying and classifying AAOCA in practical clinical settings.

Although we reported the whole result based on strategy one, our goal was to make it more generalizable and robust by utilizing all available labeled datasets for further use. The performance of the model remained consistent when tested with data from the external test set under the second strategy, and in both the second and third strategies, we did not find more positive cases from the third center; however, the number of false positive cases decreased. This demonstrates the robustness and generalizability of the model developed using the first and main strategy. However, for future applications, we recommend developing the model using the third strategy, as it utilizes the entire dataset from multiple centers and is more likely to produce generalizable results in real-world scenarios.

A recent study by Pascaner et al. [246] proposed an automated deep learning-based segmentation and detection method for AAOCA, focusing on the segmentation of the aorta and coronary arteries. They used a small, single-center dataset to develop a segmentation model based on 124 CCTA scans, reporting an accuracy, precision, and recall of 1 in a test set comprising only 13 images. Their method relies on the segmentation of coronary arteries using multi-view 2D segmentation of CCTA, followed by post-processing of the segmentation output, rule-based post-processing such as connectivity, and characteristic analysis of anomalous using a decision tree model. Although the authors implemented interesting approaches, the study has several limitations. Firstly, the small single-center dataset set for training, evaluation, and testing, without any external test set, limits the model's robustness and generalizability. Moreover, the segmentation model was trained on only 99 images, with limited variability across anomalies. Anomalies can vary significantly within datasets, and the segmentation model could easily fail on other anomalies as even the normal coronary arty segmentation is a challenging task despite the availability of a large dataset [235]. As the detection of anomalies relies on the segmentation model, any mis-segmentation could lead to failures in the classification model. In contrast, our study addresses these limitations by employing a rough segmentation of larger structures, used solely for cropping purposes. We utilized a larger training dataset to train a classification model based on a deep learning network. Moreover, we internally and externally tested the model with a large dataset, demonstrating its effectiveness in real clinical situations. Additionally, we developed two more models that classify the anomalous coronary artery and its risks. Our pipeline can be used in fully automated approaches as well as in physician-in-the-loop approaches.

The current study bears some limitations due to the low prevalence of specific anomalies; for instance, conditions like anomalous left coronary artery from the pulmonary artery (ALCAPA) are not included in our training dataset. We hypothesize that the network still recognizes such cases as anomalies based on learned patterns of normal coronary arteries. This shows the potential for future studies to use unsupervised learning or out-of-distribution analysis to better manage anomaly types not represented in the training data. Another limitation is using contrast-enhanced CTimages specifically acquired for cardiac assessments. Developing a deep learning model that detects anomalies in non-contrast images, such as standard chest CTs, would greatly expand its utility beyond cardiac imaging to include applications like any chest CTscans. However, detecting coronary arteries in non-contrast images presents significant challenges due to low image contrast and possibly low resolution, complicating the analysis even for highly experienced cardiologists and radiologists. While our model has undergone rigorous testing, including internal prospective tests, external testing datasets, and external clinical evaluation datasets to simulate real-world applications, it has yet to be implemented in a real clinical setting. To fully assess its practical utility, this model should be deployed prospectively in real-world clinical environments or assessed in randomized clinical trials [247, 248] to collect direct feedback and performance metrics.

4.5 CONCLUSION

We have developed an artificial intelligence tool designed for the fully automated detection and classification of AAOCA using CCTA images. This tool can operate seamlessly alongside clinical assessments of CCTA scans, providing real-time alerts to medical personnel about potential high-risk AAOCA anatomy, which is important in a rare disease to which physicians are unfrequently exposed. Furthermore, it can be utilized to identify AAOCA in large CCTA patient cohorts, given that the risk associated with the disease is currently unknown and warrants future investigation. Therefore, this tool could potentially enhance diagnostic efficiency, assist in managing AAOCA, and improve outcomes in these patients.

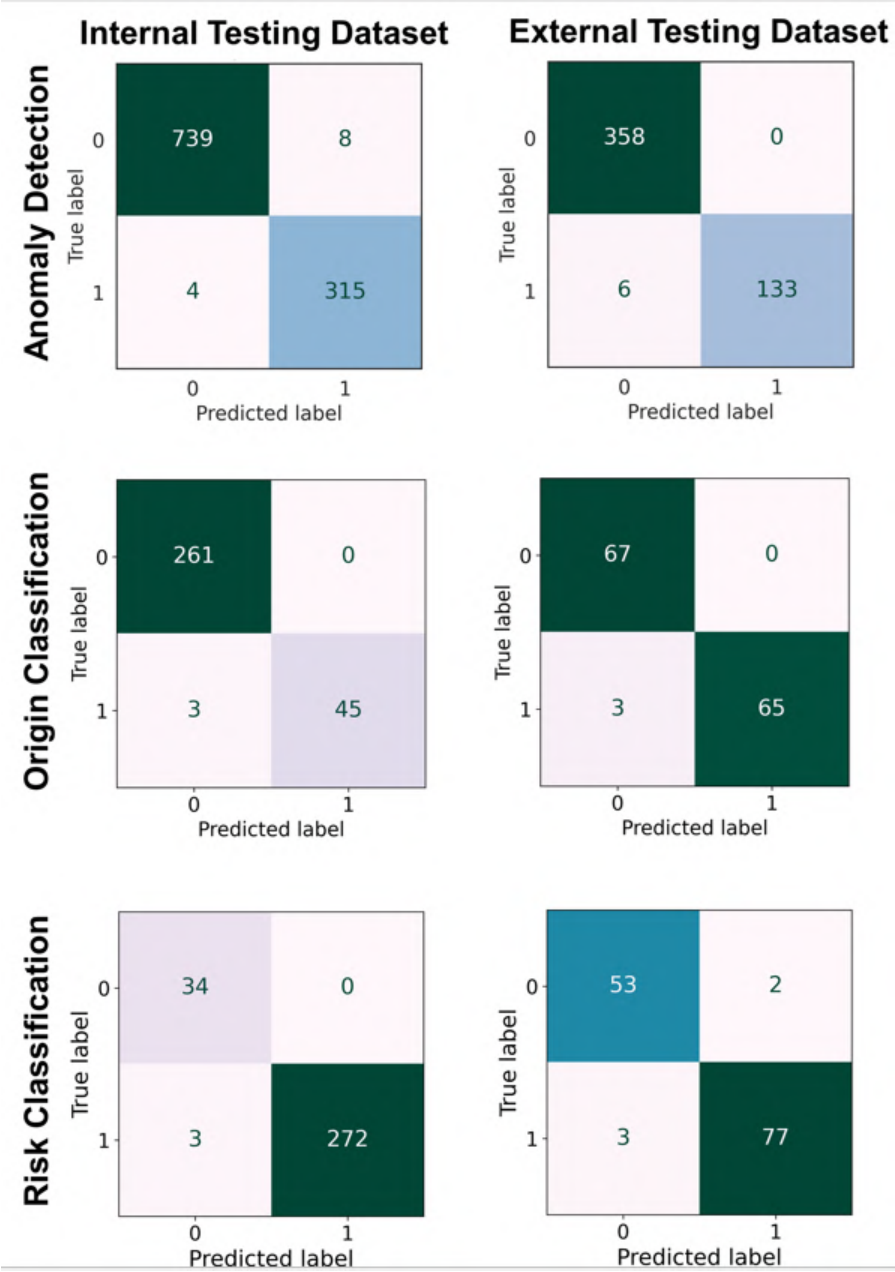


Figure 25: Confusion matrices of different models in various Tasks for different datasets in the ensemble model.

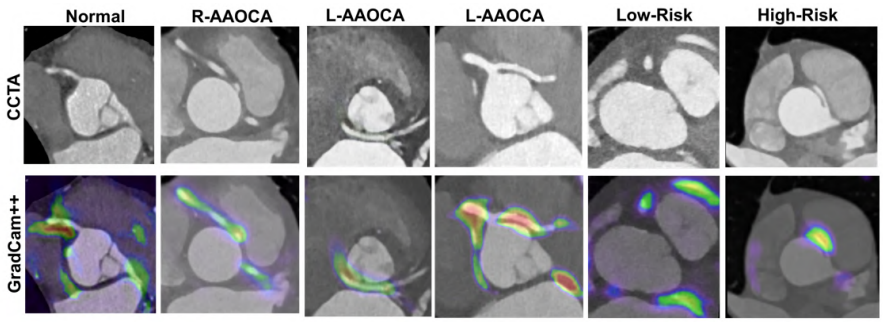


Figure 26: CCTA images and corresponding GradCam++ features for normal cases and different anomalies.

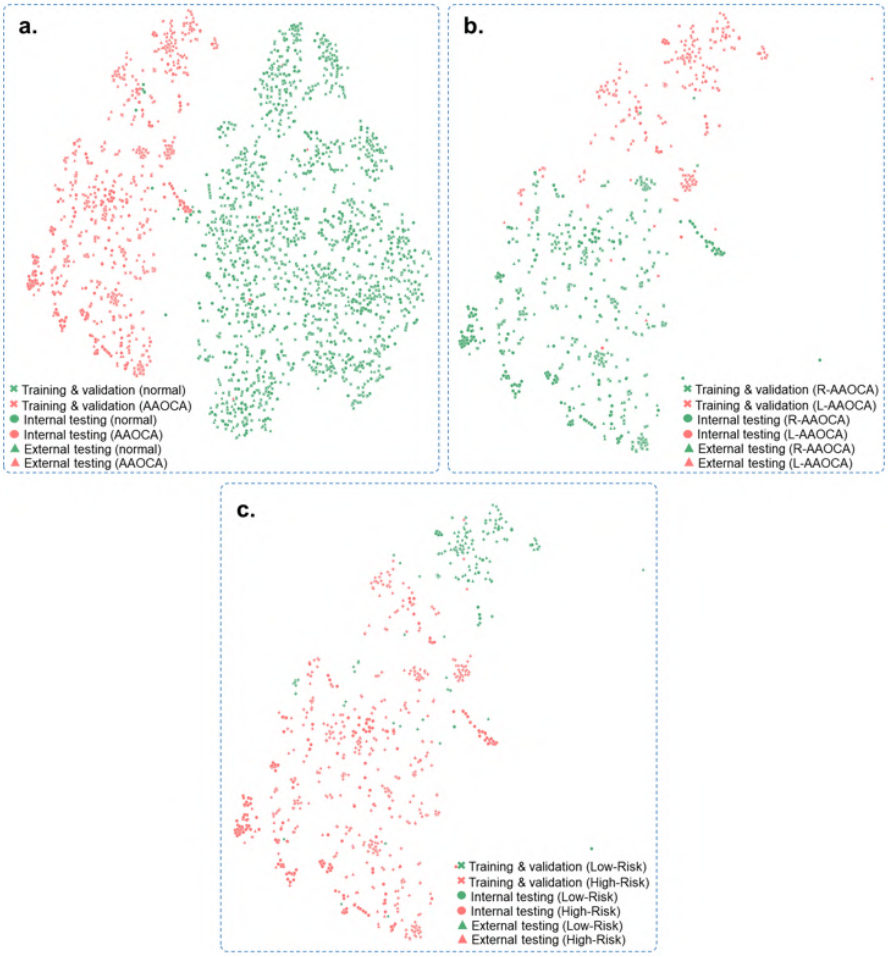


Figure 27: t-SNE maps of the Task 2 model (anomaly detection) colorized for a) anomalies and normal cases. Only for the anomaly dataset, b) right and left anomalies, and c) high and low-risk anomalies.

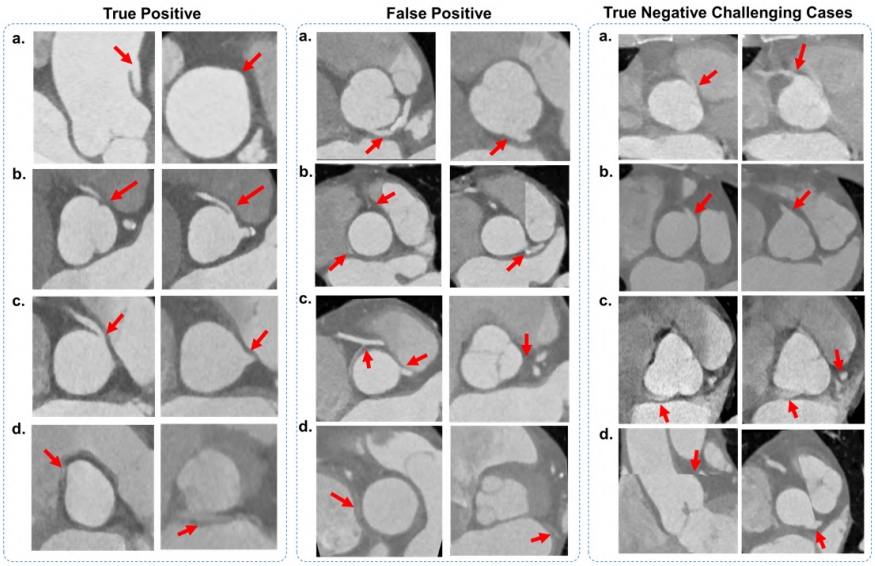


Figure 28: Developed model applied on external clinical evaluation dataset (unlabeled dataset) for real-world screening scenario. True Positives: a) High take-off of the right coronary artery (RCA) with low-risk anatomy, b, c) Right coronary artery originating from the left coronary sinus (R-AAOCA) with high-risk anatomy, d) Circumflex arteries originating from the right coronary sinus (L-AAOCA, left circumflex). False Positives: a) Left coronary artery originating very close to the non-coronary sinus, b) Appearance of a very thin coronary artery (conus artery) mimicking an R-AAOCA and contrast agent artifact mimicking L-AAOCA (left circumflex), c) left circumflex like anomaly (conus artery) with the left artery originating very close to the right sinus, along with the disappearance of the coronary sinus border in that region, d) Very faint coronary artery resembling the left circumflex (L-AAOCA) and a very thin left circumflex in the image. True Negative Challenging Cases: a, b) Motion artifact creating an RCA-AAOCA-like coronary, c) Motion artifact generating an L-AAOCA-like (left circumflex) anomaly, d) Motion artifact removing the connection of the left coronary artery to the left sinus.

5 TRANSTHYRETIN AMYLOID CARDIOMYOPATHY

As for the previous chapter, the study presented here was conducted during my visiting period at the Bern University Hospital. The work has resulted in the following publication:

- Isaac Shiri, Sebastian Balzer, **Giovanni Baj** et al. "Multi-modality artificial intelligence-based transthyretin amyloid cardiomyopathy detection in patients with severe aortic stenosis." In: *European Journal of Nuclear Medicine and Molecular Imaging* (Sept. 2024). DOI: 10.1007/s00259-024-06922-4

5.1 INTRODUCTION

Transthyretin Amyloid Cardiomyopathy (ATTR-CM) is a progressive, underdiagnosed cardiac disorder caused by the deposition of transthyretin fibrils, which leads to heart failure and worsened prognosis. It frequently coexists with Aortic Stenosis (AS), increasing the risk of adverse outcomes in patients undergoing Transcatheter Aortic Valve Implantation (TAVI). Diagnosis of ATTR-CM typically involves clinical exams, ECG, echocardiography, and cardiac imaging, but often requires invasive testing for confirmation. Developing a non-invasive, cost-effective method for early detection is crucial for improving patient outcomes. Different studies have applied AI to detect and screen for ATTR-CM across different data modalities [82, 249, 250]. AI-driven algorithms employing both DL and ML across clinical, echocardiography, ECG, scintigraphy, and CMR imaging have shown enhanced diagnostic accuracy [82]. Our study's main aim is to develop and comprehensively evaluate ML models using a pre-procedural and routinely collected TAVI multimodality dataset for detecting ATTR-CM. By evaluating ML algorithms within and across modalities in the same patient cohort, we offer insights into the strengths and limitations of each approach for ML-based ATTR-CM detection using different modalities.

5.2 MATERIALS AND METHODS

Figure 29 presents the study overview, including data collection, preprocessing, model training, validation, testing, and reporting phases utilized in the current study. The study follows the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence (TRIPOD+AI) statement [150].

5.2.1 *Study Design and Population*

The data for this study were collected from multiple modalities to enable a comprehensive analysis of ATTR-CM detection in patients with severe AS planned for TAVI [251, 252]. Consecutive patients (between August 2019 and 2021) with symptomatic severe AS in the absence of known cardiac or extra-cardiac amyloidosis were referred for TAVI at Bern University Hospital and recruited in the ATTR-AS (Amyloid Transthyretin in Aortic Stenosis, NCT04061213) study (ClinicalTrials.gov: NCT04061213) were considered eligible [251, 252]. The study design was approved by the Bern ethics committee, conducted in accordance with the Declaration of Helsinki, and study participants provided written informed consent before any data collection [251, 252]. Baseline and follow-up clinical data were prospectively recorded in a dedicated database held at the clinical trials unit of Bern University Hospital [251, 252]. These included clinical assessments, laboratory tests, ECG, and echocardiography (transthoracic echocardiography). Additionally, left and right heart catheterization and various forms of interventional imaging (integrated transesophageal echocardiography and invasive measurements were utilized. The patients underwent diagnostic evaluation with the following advanced cardiac imaging modalities: SPECT and 4D-CT. Clinical characterization of patients is provided in Table 12. Clinical follow-up involved standardized interviews, documentation from referring physicians, and hospital discharge summaries. A dedicated clinical event committee collected and adjudicated adverse events based on Valve Academic Research Consortium-2 criteria [251–253].

5.2.2 *ATTR-CM diagnosis*

As part of the ATTR-AS (NCT04061213) study, all patients underwent [99mTc]-3, 3-diphosphono-1, 2-propanodicarboxylic acid ([99mTc]-DPD) scintigraphy for ATTR-CM screening [251, 252]. Approximately 3 hours post intravenous injection of 700 ± 70 MBq [99mTc]-DPD, whole-body planar images were acquired (15 cm/min) using a dual-head hybrid SPECT/CT system (Intevo; Siemens Healthineers) equipped with low-energy high-resolution (LEHR) collimators [251, 252]. The images were reconstructed using a high-order low-pass Butterworth filter (order of 5) and a zoom of 1.0, using a 256×256 matrix size [251, 252]. Following planar imaging, a SPECT /CT scan of the thorax was carried out using a step-and-shoot method adjusted for body contour (32 steps each 30 seconds, zoom of 1.0, 256 matrix size) [251, 252]. Then, SPECT images were reconstructed using an iterative algorithm (OSEM, 4 subsets, 8 iterations), supplemented by a 12-mm Gaussian filter [251, 252]. Additionally, a low-dose CT scan was conducted for attenuation correction, using 130 kV with CareDose, a pitch of 1.2, a

Table 12: Clinical characterization of the current study patient population. More detailed parameters are presented in reference [252].

	All (n=263)	**ATTR-CM Neg (n=236)**	**ATTR-CM Pos (n=27)**	**P-Value**
		Clinical		
Sex	M: 56.7%	M: 4.6%	M: 15.4%	0.002
	F: 43.3%	F: 96.5%	F: 3.5%	
Age (y)	82.7 ± 4.6	82.4 ± 4.5	85.3 ± 4.6	0.002
BMI	26.6 ± 5.2	26.7 ± 5.2	26.0 ± 4.5	0.53
BSA (m^2)	1.85 ± 0.22	1.8 ± 0.2	1.9 ± 0.2	0.52
CAD	41.4%	40.7%	48.1%	0.46
		Intervention imaging		
AVA Echo (cm^2)	0.7 ± 0.29	0.68 ± 0.26	0.89 ± 0.43	<0.001
		Laboratory		
NT-proBNP (μg/L)	2.83 (0.56, 3.49)	1.14 (0.51, 2.95)	4.46 (1.93, 6.30)	<0.001
CREAT (mmol/L)	95.3 ± 34.2	94.1 ± 33.3	105.6 ± 40.7	0.10
		CT		
CT-AVC (kau)	2.48 (1.64, 3.75)	2.55 (1.68, 3.76)	2.21 (1.29, 3.66)	0.86
CT Days to scint.	1 (0, 1)	1 (0, 1)	1 (0, 1)	0.37
DLP (mG·cm)	973 ± 415.6	959.3 ± 422.6	1088.1 ± 336.4	0.13
Contr. agent (mL)	86.7 ± 12.7	86.8 ± 12.7	86.3 ± 13.3	0.85
		Echocardiography		
LVEF (%)	53.7 ± 12.0	54.3 ± 12.1	50.2 ± 11.2	0.36
LV mass (g)	223 ± 70.2	217 ± 70.3	267 ± 55.2	0.06
LVMi (g/m^2)	120 ± 36.9	117 ± 37.6	141 ± 24.9	0.049
LVST (mm)	13.6 ± 2.8	13.4 ± 2.8	14.5 ± 2.8	0.38
LVPWT (mm)	11.6 ± 2.2	11.4 ± 2.2	13.1 ± 1.8	0.046
		CT Global Strain (%)		
LV GLS	-14.3 ± 4.7	-14.7 ± 4.8	-11.2 ± 3.2	<0.001
LV GRS	49.5 ± 25.4	50.2 ± 25.8	42.9 ± 21.0	0.16
LV GCS	-17.6 ± 6.3	-17.9 ± 6.4	-15.3 ± 5.2	0.04
RV GLS	-18.4 ± 7.2	-18.7 ± 7.2	-16.2 ± 6.9	0.09
LA GLS	14.2 ± 9.9	14.9 ± 10	8.3 ± 7.2	0.001

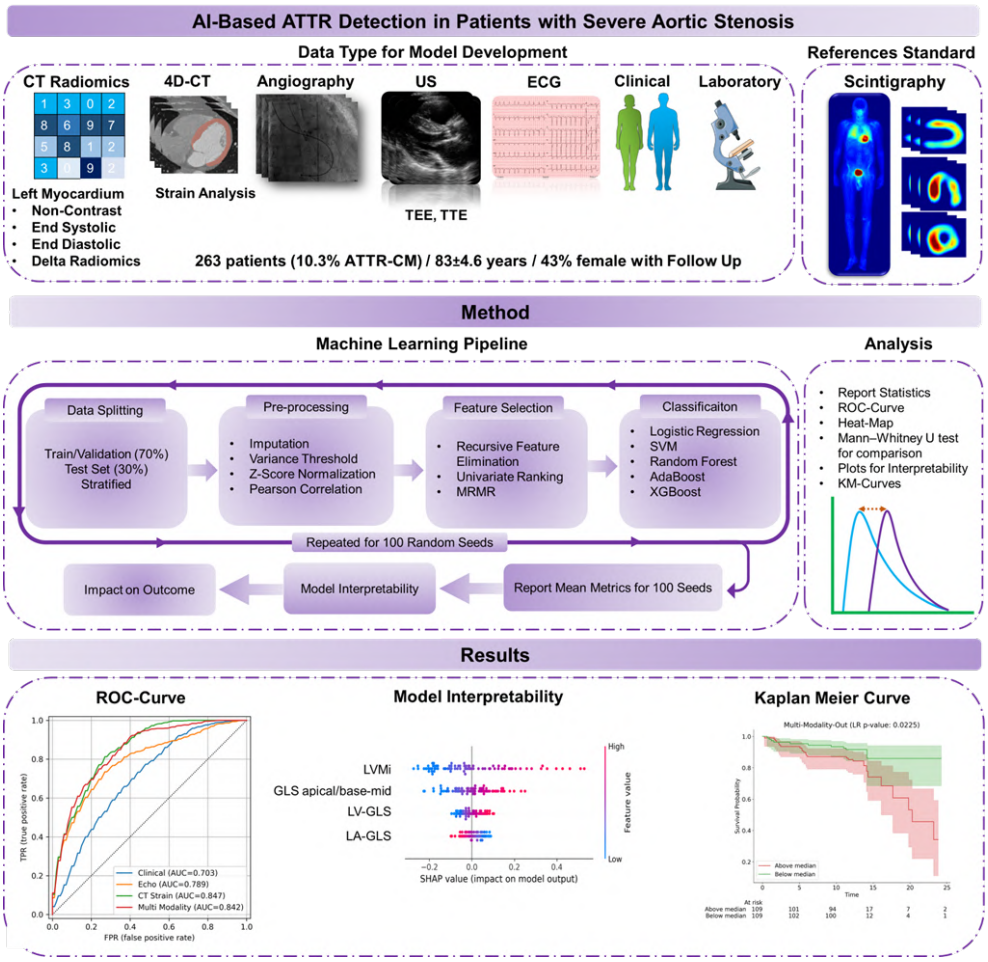


Figure 29: The flowchart of the current study with the study design, including the phases of data collection, preprocessing, model training, and validation.

rotation time of 0.6 seconds, and a collimation of 16×0.6 [251, 252]. The [99mTc]-DPD scintigraphy results were interpreted as positive for participants exhibiting moderate to high myocardial tracer uptake (Perugini grade 2 or 3) and negative for those with no or low uptake (Perugini grade 0 or 1), as assessed by nuclear medicine physicians and cardiac imaging cardiologist all with >10 years of experience in nuclear cardiology [251, 252]. More information on data was provided in [251, 252].

5.2.3 *Data Preparation and Image Processing*

Advanced image processing techniques were utilized to extract CT strain and LV mass and function information from 4D Contrast-Enhanced Computed Tomography (CECT) images; more details on the acquisition and processing of 4D CECT were previously published [251, 252]. Radiomics features were extracted from the LV myocardium us-

ing CT images, including non-contrast images and images from the diastolic and systolic phases of contrast-enhanced 3D-CT. Delta radiomics were calculated using the diastolic and systolic phases of CECT. Segmentation of the LV was initially provided by an automatic approach and subsequently evaluated and modified as needed for different images. Various radiomics features, including intensity, shape, and second and higher-order features, were extracted using the Image Biomarker Standardization Initiative (IBSI) [254, 255] consensus Python library [256] with bin discretization set to 64 and an isotropic voxel size of 1 mm^3. We extracted the following number of features from each modality: 101 features from clinical, 13 from laboratory, 18 from ECG, 34 from echocardiography, 34 from invasive measurements, 6 from interventional imaging, 76 from CT strain, and 420 from radiomics (end-systolic, end-diastolic, non-contrast, and delta phases). The dataset was initially split into a training (70%) and a hold-out test set (30%) with stratification regarding the ATTR-CM status. Missing data within the dataset was imputed using an iterative imputation technique (applying a round-robin approach).

5.2.4 *Machine-learning Algorithm*

Z-score normalization was applied to the features to ensure uniformity in scale. Features exhibiting low variance (below a threshold of 0.99) were discarded. Subsequently, features with high correlation (Pearson correlation coefficient higher than 0.95) were grouped, and only the most predictive feature from each group was retained. Following this initial preprocessing, various feature selection algorithms, including Recursive Feature Elimination (RFE), Univariate Ranking (UniVa), and Minimum Redundancy Maximum Relevance (MRMR), were employed on the feature set to select informative features. Using the selected features, a variety of classifier models were trained, including Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), AdaBoost, and eXtreme Gradient Boosting (XGB).

5.2.5 *Parameters and Hyperparameters Optimization*

All ML model development steps, including preprocessing and feature selection, were performed exclusively on the training set and validation (70%) to process and select important features. Using these selected features, classifier parameters and hyperparameters were optimized using grid search on the training set to build the optimized model. Subsequently, the developed models were evaluated on the hold-out test set. This approach ensured that there was no possibility of information leakage between the training and test sets. The entire process (from data splitting to model evaluation) was repeated 100 times with a random seed to assess the robustness of the models.

5.2.6 *Evaluation and Statistics*

The SHapley Additive exPlanations (SHAP) model is utilized to interpret the outputs of ML models, providing insights into the contribution of each feature to the ATTR-CM prediction. The performance of these models was assessed using different metrics, including balanced accuracy (sensitivity+specificity)/2), receiver operating characteristic area under the curve (ROC-AUC), sensitivity, and specificity. The Mann–Whitney U test was employed to evaluate differences in performance metrics for statistical comparison across different models and modalities. Additionally, cumulative incidence curves were plotted for the diagnostic model output and their features to visualize the impact on patient outcomes over time. All ML was implemented by `Scikit-learn` in the Python programming language [133], and all models and code are publicly available in the GitHub repository (`https://github.com/AI-in-Cardiovascular-Medicine/ML_pipeline_tabular`).

5.3 RESULTS

5.3.1 *Study Population*

Out of 489 patients initially assessed, 91 were ineligible, and 83 did not consent [251, 252]. Thus, 315 patients consented and were enrolled [251, 252]. From this cohort, 51 were excluded due to the absence of 4D-CECT, and one was excluded due to lack of correct image phases [251, 252]. Finally, 263 patients (83 ± 4.6 years, 114 females) who underwent [99mTc]-DPD scintigraphy and had available data from multiple modalities, including 4D-CECT, were included in the analysis [251, 252]. ATTR-CM was confirmed in 27 (10.3%) of these patients [251, 252]. Among those diagnosed with ATTR-CM, 22 underwent genetic testing, which revealed that 21 (95%) had wild-type ATTR, and 1 (5%) exhibited a transthyretin mutation (Val40Met) [251, 252]. The mean (standard deviation) follow-up for all causes of mortality and cardiovascular mortality was 13 (5) months, and it was available for 218 patients (23 ATTR-CM).

5.3.2 *Diagnostic Performance*

Figure 30 summarizes the performance of the diagnostic modalities (the best-performing ML algorithm in each modality) evaluated through different metrics.

Diagnostic performance of conventional first-line diagnostic tests

ECG data showed a low performance with a mean (standard deviation) ROC-AUC of 0.67 (0.08), sensitivity of 0.45 (0.16), and specificity of 0.89 (0.04). Clinical data demonstrated a moderate discriminatory performance with ROC-AUC of 0.70 (0.09), sensitivity of 0.81 (0.18), and specificity of 0.63 (0.18). The laboratory data showed moderate results, with a ROC-AUC value of 0.76 (0.07), sensitivity of 0.82 (0.13), and specificity of 0.72 (0.11). Echocardiography resulted in acceptable performance with ROC-AUC of 0.79 (0.09), sensitivity of 0.80 (0.15), and specificity of 0.78 (0.11).

Diagnostic performance of the CT Radiomics

In the CT radiomics, CT non-contrast radiomics showed the lowest performance with a mean (standard deviation) ROC-AUC of 0.68 (0.11), sensitivity of 0.76 (0.19), and specificity of 0.66 (0.21). The performances of CT diastolic, CT systolic, and CT delta radiomics were similar, with no statistically significant differences between the metrics of diagnostic performance (p-value > 0.05). Although the combined evaluation of all CT radiomics features resulted in a mean (standard deviation) of ROC-AUC of 0.74 (0.07), sensitivity of 0.81 (0.14), and specificity of 0.68 (0.13), it did not outperform the individual systolic and diastolic radiomics (p-value > 0.05).

Diagnostic performance of the CT Strain

The highest discriminatory performances were observed for CT strain, yielding the highest diagnostic accuracy compared to other modalities. CT strain achieved the highest ROC-AUC of 0.85 (0.05), sensitivity of 0.90 (0.11), and specificity of 0.74 (0.11).

Diagnostic performance of invasive modalities

The lowest performances were observed for Invasive Cath (ROC-AUC of 0.61 (0.09), sensitivity of 0.64 (0.25), and specificity of 0.67 (0.28)). Interventional Imaging showed a ROC-AUC of 0.70 (0.08), sensitivity of 0.79 (0.18), and specificity of 0.63 (0.19).

Diagnostic performance of the multi-modality

The multi-modality approach in which features from different diagnostic modalities were jointly considered, yielded a high mean (standard deviation) ROC-AUC of 0.84 (0.06), sensitivity of 0.87 (0.13), and specificity of 0.76 (0.12). This is comparable to CT strain, indicating that while combining features can achieve high diagnostic performance, it does not statistically significantly outperform CT strain.

The best model and modality for identifying ATTR-CM in patients with severe AS

Between different modalities, four models from CT strain (Manual+LR, RFE+LR, UniVa+AdaBo, and MRMR+SVM) and one model from multi-modality (RFE+LR) showed the highest performances, with no statistically significant differences between these models.

ROC and heatmap plots in different models and modalities

Figures 31 and 32 present the ROC curves of the top three models across different modalities. Figure 33 shows heatmaps of different feature selection and classifier combinations for different metrics within the echocardiography, CT strain, and multi-modality model.

5.3.3 *Interpretability*

Figure 34 presents the SHAP summary of the top three models, including echocardiography, CT strain, and multi-modality. The SHAP plots illustrate the relative importance of each feature in each model, providing insights into the decision-making processes underlying the feature-outcome relationships in each modality toward ML model interpretability. For instance, in the echocardiography model, features such as the mean gradient of the aortic valve and maximum septal wall thickness had the highest impact. Decreasing the mean gradient of the aortic valve and increasing the maximum septal wall thickness positively influenced the ML model's output toward ATTR-CM positive diagnosis. In the CT strain model, increasing feature values like LV global longitudinal strain apical /base and mid, LV global longitudinal strain (LV-GLS, %), end-diastolic LV mass index (LVMi, g/m2) and decreasing in left atrial global longitudinal strain (LA-GLS, %) has a positive influence on model output. In the multi-modality model, in addition to the CT strain and echocardiographic features, other variables such as age, radiomics features, and laboratory data contributed to the model's outcomes.

5.3.4 *Prognostication information of diagnostic features*

Prognostication based on ATTR-CM status did not show significant prognostic information (indicated by a log-rank p-value > 0.05). Among the different diagnostic ML models, only the multi-modality model provided prognostic information for distinguishing between low- and high-risk groups for all-cause mortality (Figure 35).

5.4 DISCUSSION

In the current study, we comprehensively evaluated the performance of ML approaches based on preprocedural and routinely collected data from different contemporary diagnostic modalities to predict ATTR-CM in patients with severe AS planned for TAVI. These modalities included clinical assessment, ECG and echocardiography, as well as more advanced imaging processing techniques such as CT strain and CT radiomics. We employed a wide range of standardized ML algorithms to predict ATTR-CM using single and multi-modality data. While echocardiography demonstrated good performance, CT strain exceeded it in accuracy. The multi-modality model did not outperform the CT strain-only data. Specific features from different modalities provided prognostic information in severe AS patients for all-cause and cardiac-specific mortality. The presence of ATTR-CM was not shown to be an outcome predictor. Current diagnostic standards for establishing an ATTR-CM diagnosis, such as biopsy or bone scintigraphy [89, 90], introduce additional costs and burdens, and are not included in the standard clinical practice for patients with severe AS undergoing TAVI procedures [82–84]. Although scintigraphy is becoming the gold standard for ATTR-CM detection and provides prognostic information [257–259], it presents other challenges, such as radiation exposure and delays in diagnosis due to the low deployment of nuclear medicine centers. Moreover, scintigraphy imaging can yield negative results if amyloid deposition is minimal at the time of examination [82–84]. Therefore, ATTR-CM is likely underdiagnosed in this cohort, [82–84] highlighting the need for non-invasive and cost-effective diagnostic tools [82]. Our study presents new evidence that utilizing ML to integrate preprocedural and routinely collected data aids in detecting concomitant ATTR-CM in patients with severe AS. Using the existing data for ATTR-CM detection could also enhance prognostication in this patient group.

Several studies have been employed to detect ATTR-CM using AI in different modalities [82, 260–262]. Previous studies using the echocardiographic images and DL model [260] and handcrafted echocardiographic parameters and ML [262] reported an average AUC of 0.87 (5-fold Cross-validation (CV)) and 0.82 (0.95, 0.76, 0.78, and 0.80 on the four external tests) in detecting cardiac amyloidosis and wild-type ATTR-CM, respectively. Another study [261] employed automated tools using ECG and echocardiography to detect wild-type ATTR-CM. The DL was externally tested at four centers, with 441 (AUC: 0.91), 369(AUC: 0.89), 229 (AUC: 1.0), and 239 (AUC: 0.96) patients [261]. In our study, using only the ECG modality did not yield good performance, whereas the echocardiographic modality provided results comparable to previous studies. Despite our limited dataset, focusing on patients with severe AS and overlapping symptoms made con-

structing high-performance models challenging. Thus, this should be considered when comparing study results, as model performance is influenced by the specific cohort used for development and evaluation, not just the metrics [247]. In a previous study [263], [99mTc]-HMDP scintigraphy was utilized to detect ATTR-CM. They [263] focused on classifying ATTR-CM based on Perugini grades using DL models. They [263] reported an AUC of 0.87 for multiclass classification, 0.94 for the binary comparison of grade < 2 vs. grade ≥ 2, and 0.89 for grade < 3 vs. grade 3 using a 5-fold CV. Another study [264] developed a DL model using scintigraphy images ([99mTc]-DPD/[99mTc]-HMDP) to detect a Perugini grade of ≥ 2 in ATTR-CM patients. They reported an AUC of 0.99 in both the development and external test phases, demonstrating high diagnostic accuracy. In our study, we did not use scintigraphy as an input for detecting ATTR-CM but rather as the ground truth, with most positive cases confirmed by pathology and genetic tests. As scintigraphy is not routinely implemented in clinical practice among patients with severe AS undergoing TAVI due to additional cost, our study demonstrates the feasibility of detecting ATTR-CM in this cohort using preprocedural and routinely collected data with good performance. This model could be used for initial screening with available data for this cohort, allowing suspected cases to undergo scintigraphy for confirmation. The diagnosis performance of ML and DL was investigated [265] using CMR images, and they reported an AUC of 0.98 for DL and 0.95 for ML for ATTR-CM detection. Other [266] conducted a study using CMR sequences to diagnose cardiac amyloidosis automatically. They [266] employed binary classification approaches to analyze single 2D slices using DL and used averaged voting across all slices for comprehensive patient-wise analysis. They reported AUC scores of 0.96 for LGE, 0.93 for MOLLI, and 0.91 for CINE. In our study, we did not use the CMR dataset due to its limited availability for TAVI patients. However, future studies could integrate this modality to evaluate new model performance.[260] In a recent study [267] contrast-enhanced CT radiomics features of 30 patients were used to differentiate cardiac amyloidosis from severe AS. Due to the small data size, they used a leave-one-out CV and reported accuracy, sensitivity, and specificity of 0.93. In another study [268] CT radiomics features were evaluated for detecting cardiac amyloidosis in AS patients who underwent TAVI. Using a 7-fold CV, they reported an AUC of 0.92 for radiomics and 0.96 when combining radiomics with clinical information.

Our study evaluated the performance of various CT radiomics features, achieving a moderate AUC of 0.75. Compared to previous radiomics studies, our dataset was larger, and we enhanced the reliability of our results by repeating the entire process 100 times with random seeds to avoid any bias in the chosen test set. This approach is essential in ML studies with small to medium-sized datasets be-

cause achieving high performance in a single repetition could be potentially due to a random split that favors easier cases in the test set, which may not realistically reflect real-world scenarios. In each ML model, we used SHAP analysis to understand the top model's decision-making. We observed that these features and decisions align with previous clinical findings, which makes the model more rational and reliable [82]. In the Echo, the ML model showed that decisions for detecting ATTR-CM are based on a combination of decreasing gradient, increasing wall thickness, and increasing LV mass and volume. The selected features and their behavior align with clinical symptoms in ATTR-CM patients, as the amyloid fibrils lead to an increased myocardial thickness, consequently decreasing the LV stroke volume [82, 87, 269]. While these features alone cannot fully represent ATTR-CM, their combined effects and the varying weights assigned to them could form a robust diagnostic model. While wall thickness, especially maximum septal wall thickness, was a key feature in the echocardiography model, it did not contribute significantly in the multimodality model, where other features were more influential. A recent study suggested [270] that wall thickness is not correlated with ATTR-CM, which aligns with our finding that in the presence of other features, such as CT strain, the importance of wall thickness decreases. In the CT strain modality ML modeling, we used automated and manual feature selection based on our previous study [252]. However, there was no statistically significant difference between the manually selected and the automated ones, and we achieved the highest performance using CT strain analysis for different modalities. Previously [252], we employed conventional standard statistical methods to evaluate models, reporting different cutoffs in AUC of 0.89 with internal bootstrapping (sensitivity of 0.96 and 0.77, and specificity of 0.58 and 0.85). However, in the current study using standardized ML approaches, we achieved an ROC-AUC of 0.85 ± 0.05 with a sensitivity of 0.90 ± 0.11 and a specificity of 0.74 ± 0.11. Although the AUC is slightly lower than the previous study [252], the ML model improved the performance of ATTR-CM detection by considering both sensitivity and specificity. High sensitivity and low specificity could be impractical in clinical settings due to the high number of false positives; thus, a model that simultaneously minimizes false positives and false negatives is preferable. Additionally, conventional statistical models often risk information leakage through internal bootstrapping and data splitting, potentially inflating performance metrics. In contrast, our standardized ML development approach avoided any information leakage, leading to more realistic and superior performance (considering both sensitivity and specificity simultaneously) with CT strain compared to our previous studies [252]. Our CT strain models demonstrated that combining features indicative of myocardial contractility and wall motion abnormalities could create a high-performance predictive model for detecting

ATTR-CM. Although CT strain may not be routinely collected, our analysis indicates that the top contributing features in the multimodality model are derived from CT strain. Attempts to build a model using only routinely collected data did not yield satisfactory performance, showing the importance of CT strain in accurate diagnosis of ATTR-CM. Considering advancements in CT scanner technology, which significantly reduce acquisition time and radiation dose [271], as well as the necessity of pre-TAVI CT images and previous guideline [272] recommendation, 4D-CT could potentially be acquired routinely in the future. This model could seamlessly be integrated into clinical routine, providing an additional tool that uses available information to identify and alert clinicians to high-risk patients for ATTR-CM as it might change the clinical decision for TAVI versus surgical therapy. Although the multi-modality model did not outperform the CT strain-only models, it incorporates additional features such as the radiomics of the LV myocardium, where an increase in value tends to indicate a diagnosis of ATTR-CM. Moreover, other features like age/troponin and creatine kinase (CK) were found to have positive and negative impacts on the model's output, respectively. Additionally, selected features from various modalities were useful in differentiating between low and high-risk groups for all-cause and cardiac-related mortality. Although overall prognostic assessments based on ATTR-CM did not show significant performance, the multi-modality model output was effective in distinguishing between different mortality risk groups. This shows the potential benefits of integrating multiple diagnostic modalities to enhance the accuracy of prognostic assessments and provide new biomarkers. A DL model developed [273] for ATTR-CM detection in scintigraphy, demonstrated that the outputs of the diagnostic models could serve as markers for prognosis and discriminate between high and low-risk groups for overall mortality. Another study [83] showed that a diagnostic model using ECG in severe AS patients undergoing TAVI indicated that the diagnostic DL model's output could predict all-cause mortality, major adverse cardiac events, and hospitalization due to heart failure. Our results align with these earlier studies, demonstrating the potential of diagnostic models to prognosticate and offer new biomarkers. The Kaplan-Meier curves are plotted based on the output of the models, stratified by the median value of the diagnostic model for ATTR-CM detection. We hypothesize that false positive cases, which may include patients with severe conditions resembling ATTR-CM, lead to worse outcomes due to impaired cardiac function. This suggests that our model may capture additional prognostic information not accounted for in the binary classification of ATTR-CM from different features of various modalities. Future studies should evaluate the prognostication performance of this model on TAVI patient cohorts.

In this study, we implemented multiple ML algorithms that yielded different performance results, which may arise from the specific characteristics of each model. LR often outperformed other models in different modalities, making it advantageous for clinical use due to its simplicity and greater interpretability. However, in some cases, such as with CT strain, which is the best modality for ATTR-CM detection, models like AdaBoost, SVM, and RF performed similarly well. The lower performance of complex models like XGB could be due to overfitting in the training set, given their high number of parameters. By using an untouched test set for evaluation, we ensured an unbiased comparison and selected the most reliable performing model. Gathering a comprehensive dataset encompassing multiple modalities for assessing ATTR-CM in AS is highly challenging, as the concordance of ATTR-CM is not routinely evaluated in clinical practice. Although our dataset may be considered medium-sized compared to previous clinical ML studies in ATTR-CM [260, 265–268], the methodology we applied ensures that the generated results are robust and repeatable. Furthermore, the clinical objective of our study is not to replace scintigraphy with our ML model but to use the model to detect potential cases, which can then be confirmed through scintigraphy. This approach has the potential to enhance clinical workflow for AS patients undergoing TAVI procedures.

One of the main limitations of this study is its reliance on a single-center and unbalanced dataset. However, we employed various approaches, including stratified splitting, avoiding any information leakage between the training and testing sets, and repeating the experiment with a random seed to provide more realistic and robust results. Although we tried data augmentation techniques such as SMOTE in the training set, it did not improve the model's performance, and we continued with the original data. Moreover, we have made our code and model publicly available to support open-source practices and the reproducibility of the study. Domain shift in ML studies can occur due to variations in data acquisition methods (i.e. changes in the scanner), population characteristics, and changes over time. These shifts can impact model performance and should be carefully monitored, even in single-center studies. Future studies should evaluate and validate our models' performance in larger, prospective, and external datasets.

5.5 CONCLUSION

In the current study, we implemented ML to evaluate the efficacy of various modalities for predicting ATTR-CM in patients with severe AS undergoing TAVI. While echocardiography, CT strains, and multi-modality demonstrated high diagnostic performance, with CT strain being the highest-performing modality, the multi-modality model did

not outperform CT strain alone. Other modalities, including LV radiometric features on CT scans, showed moderate performance for detecting ATTR-CM. Moreover, some diagnostic features could provide more insights for prognostication in severe AS. Our study demonstrates that applying ML to routine pre-TAVI data can effectively detect concomitant ATTR-CM in patients with severe AS, presenting a potential alternative to scintigraphy or invasive biopsies.

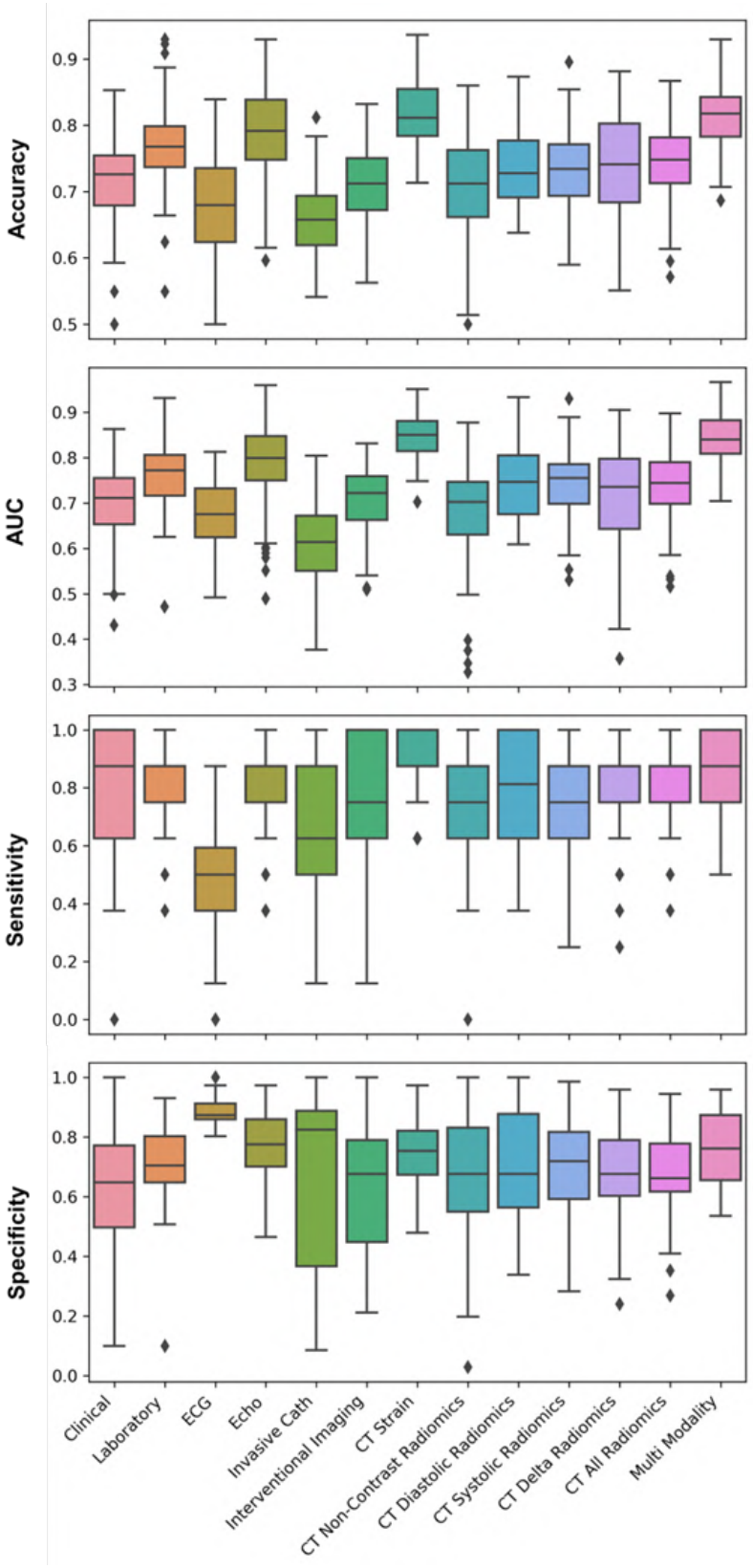


Figure 30: Comparative analysis of different metrics, including Accuracy, AUC, Sensitivity, and Specificity for the best-performing models in each modality, evaluated across 100 iterations. Clinical: RFE+LR, Laboratory: UniVa+LR, ECG: RFE+AdaBoost, Echo: UniVa+SVM, Invasive Cath: MRMR+LR, Interventional Imaging: UniVa+LR, CT Non-Contrast Radiomics: RFE+LR, CT Diastolic Radiomics: UniVa+LR, CT Systolic Radiomics: UniVa+LR, CT Delta Radiomics: UniVa+LR, CT All Radiomics: UniVa+LR, CT Strain: RFE+LR, Multi-Modality: RFE+LR.

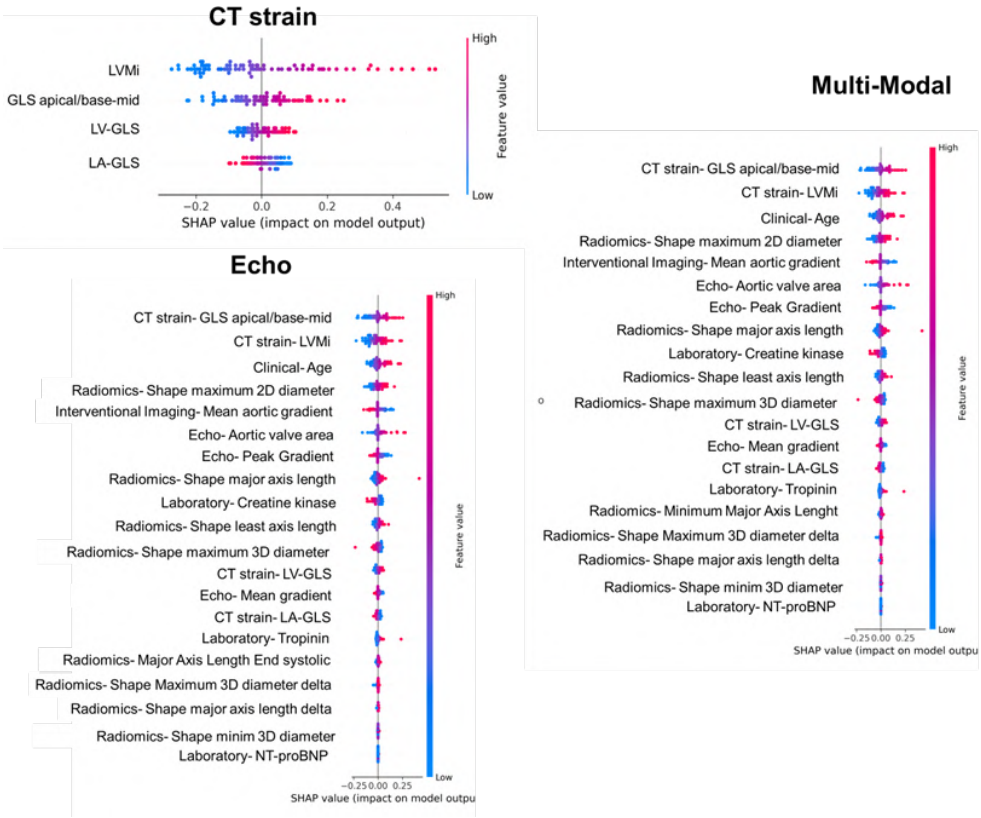


Figure 31: ROC curve of best-performing models in each modality. Clinical: Strat. 1 (RFE+LR), Strat. 2 (UniVa+LR), Strat. 3 (MRMR+LR); Laboratory: Strat. 1 (RFE+LR), Strat. 2 (UniVa+LR), Strat. 3 (MRMR+LR); ECG: Strat. 1 (RFE+AdaBo), Strat. 2 (UniVa+ AdaBo), Strat. 3 (MRMR+ AdaBo); Echo: Strat. 1 (RFE+LR), Strat. 2 (UniVa+SVM), Strat. 3 (MRMR+LR); Invasive Cath: Strat. 1 (RFE+AdaBo), Strat. 2 (UniVa+LR), Strat. 3 (MRMR+LR); Interventional Imaging: Strat. 1 (RFE+LR), Strat. 2 (UniVa+LR), Strat. 3 (MRMR+LR); CT Non-Contrast Radiomics: Strat. 1 (RFE+LR), Strat. 2 (UniVa+LR), Strat. 3 (MRMR+LR); CT Diastolic Radiomics: Strat. 1 (RFE+LR), Strat. 2 (UniVa+LR), Strat. 3 (MRMR+LR); CT Systolic Radiomics: Strat. 1 (RFE+LR), Strat. 2 (UniVa+LR), Strat. 3 (MRMR+LR).

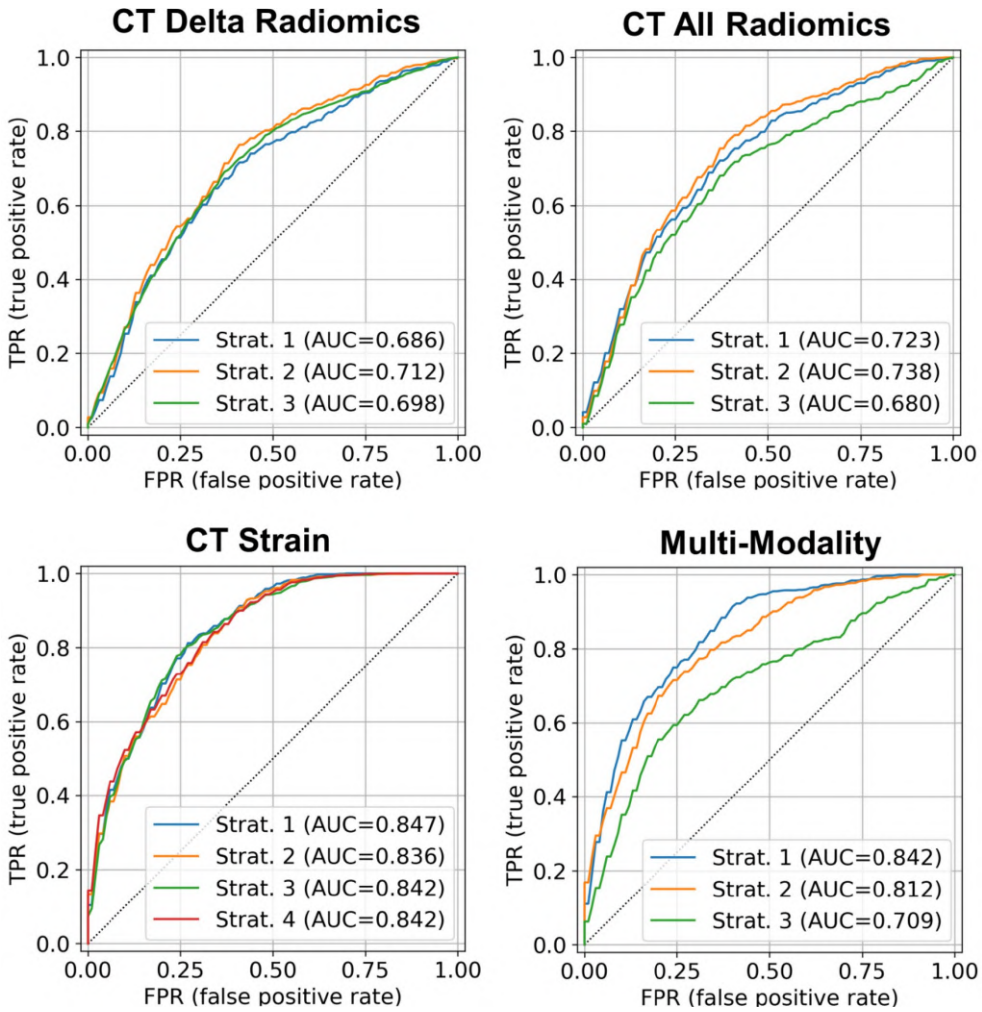


Figure 32: ROC curve of best-performing models in each modality. CT Delta Radiomics: Strat. 1 (RFE+LR), Strat. 2 (UniVa+LR), Strat. 3 (MRMR+AdaBo); CT All Radiomics: Strat. 1 (RFE+LR), Strat. 2 (UniVa+LR), Strat. 3 (MRMR+LR); CT Strain: Strat. 1 (Manual+LR), Strat. 2 (RFE+LR), Strat. 3 (UniVa+AdaBo) Strat. 4 (MRMR+SVM); Multi-Modality: Strat. 1 (RFE+LR), Strat. 2 (UniVa+LR), Strat. 3 (MRMR+LR).

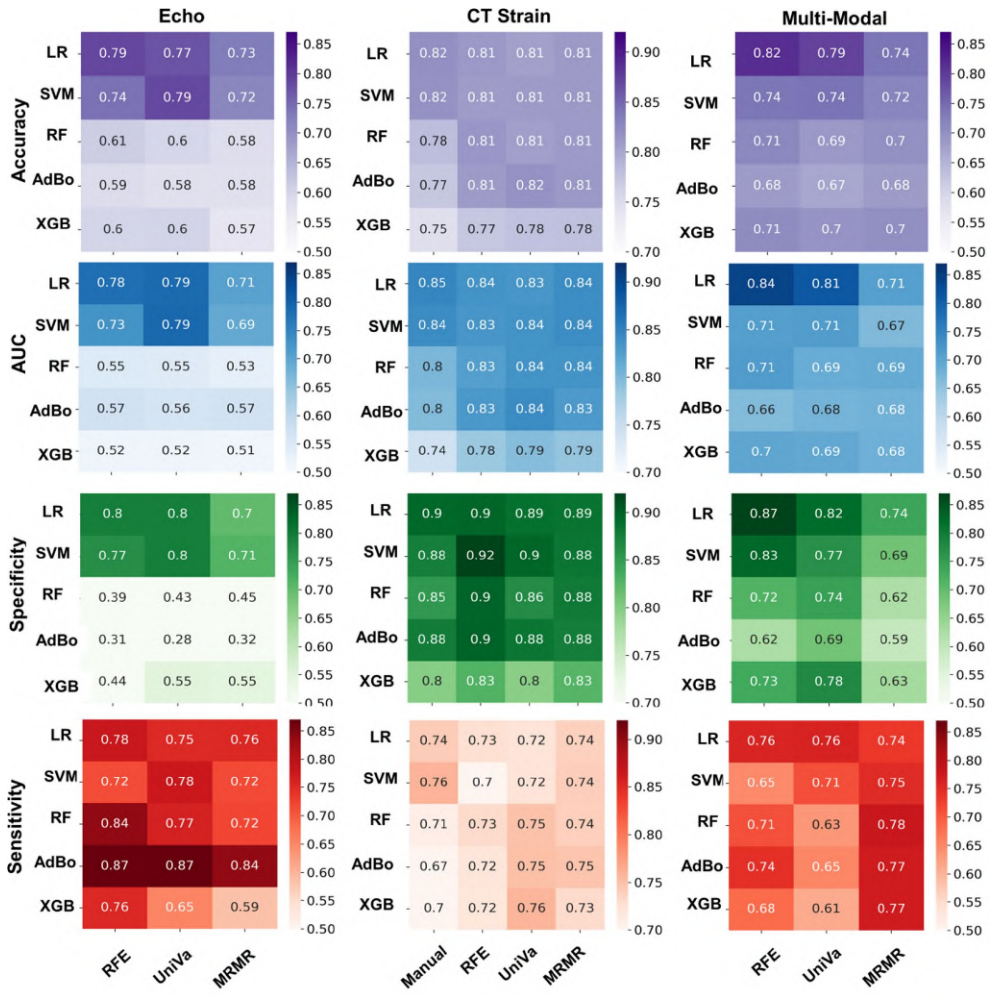


Figure 33: Heat maps displaying various metrics for echocardiography (Echo), CT strain, and Multi-Modal data.

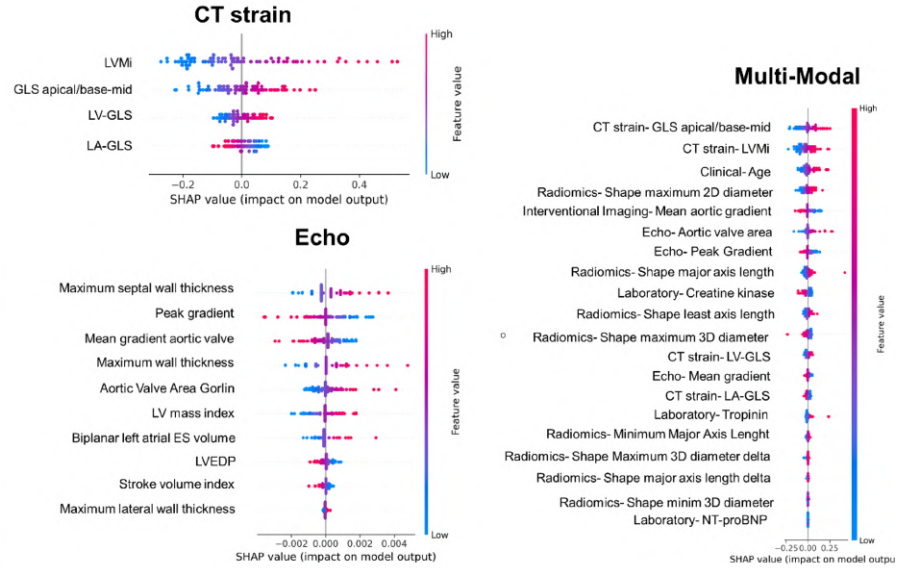


Figure 34: SHAP summary plot displaying the impact of various features across Echo, CT Strain, and Multi-Modality models. This visualization highlights the contribution of individual features to each model's predictive performance for ATTR-CM detection. LVMi: Left ventricular mass (end-diastolic) index to gr/m2, GLS apical/base-mid: Left ventricular global longitudinal strain apical /base and mid, LV-GLS: Left ventricular global longitudinal strain (%), LA-GLS: Left atrial Global longitudinal strain (%), Mean gradient aortic valve: Mean gradient aortic valve [mmHg], Maximum Septal Wall Thickness: maximum septal wall thickness of the left ventricle [mm], Peak Gradient: peak pressure gradient of the aortic valve [mmHg], Mean Gradient: mean pressure gradient of the aortic valve [mmHg], Biplanar Left Atrial ES volume: Biplanar LAESVi [ml / BSA in m2], LV mass index: left ventricular mass index LVMi [g/m2], LVEDP: estimated left ventricular end-diastolic pressure [mmHg], Maximum lateral wall thickness: maximum lateral wall thickness of the left ventricle [mm], NT-proBNP: N-terminal pro-B-type natriuretic peptide [pg/ml]

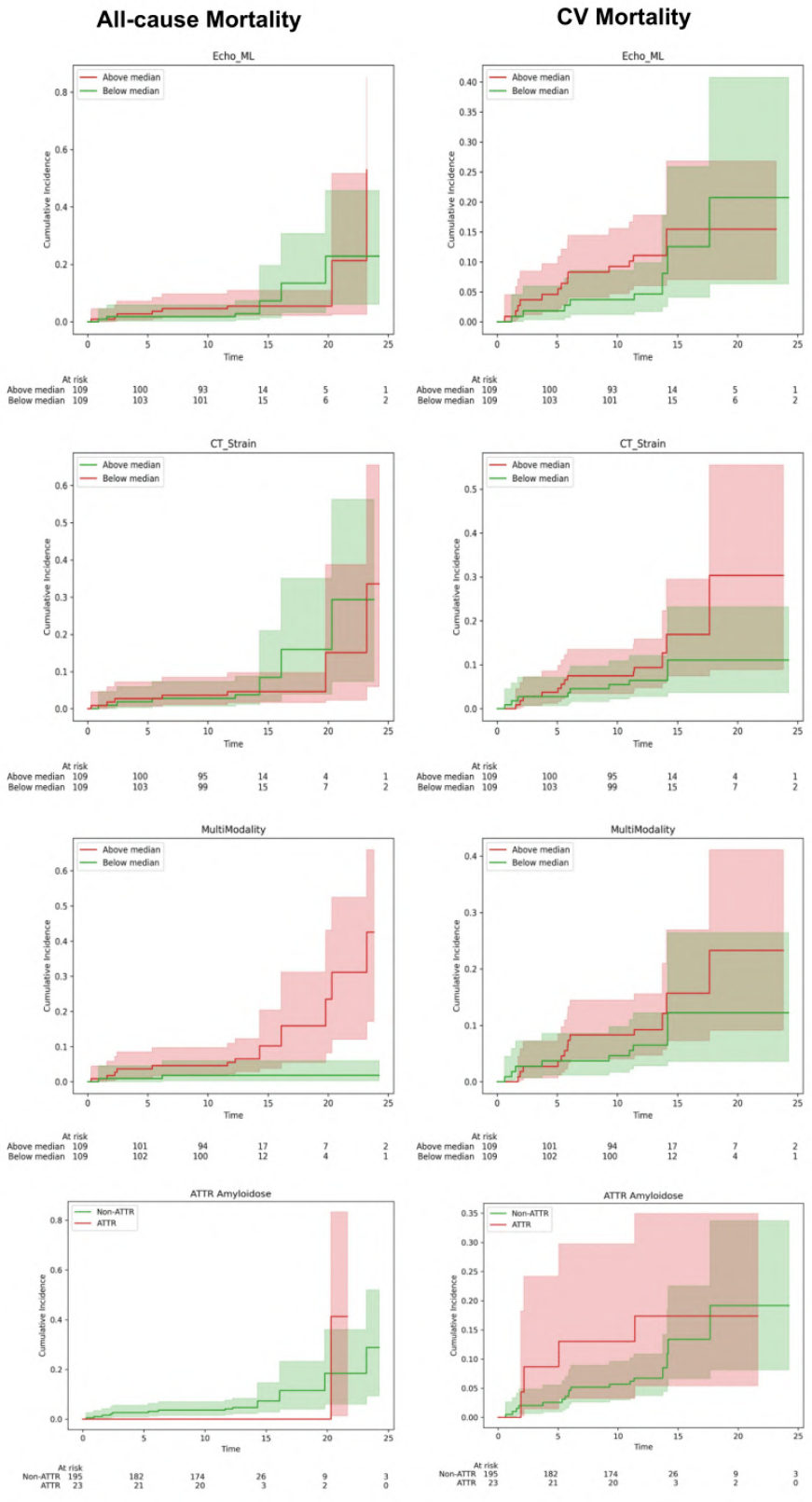


Figure 35: Cumulative-incidence curves (Aalen–Johansen estimator) for all-cause mortality and cardiovascular mortality, stratified based on the median value of the outputs of the diagnostic model for ATTR-CM and the ground truth of ATTR-CM. The stratification categories are above and below the median ML output values, illustrating survival probabilities over time for each group with the x-axis in months.

6

CONCLUSIONS

Cardiovascular Diseases are the leading cause of mortality globally, responsible for an estimated 18.6 million deaths annually. As populations age, the incidence of CVD is projected to increase significantly, and the economic burden is also substantial, with CVD-related costs expected to rise to $1.1 trillion in the U.S. by 2035. AI has the potential to enhance cardiovascular care in several key areas, from diagnosis and treatment to outcome prediction. Machine Learning and Deep Learning techniques have shown promise in analyzing complex medical data, matching or surpassing physician performance in many detection and prediction tasks. Despite progress, AI technologies have to face many challenges and their integration into clinical practice remains in its infancy. The first requirement for this to happen is that models have to be developed and evaluated with rigorous methodologies and within a robust statistical framework [274]. It is in light of this that the contributions of this thesis have been developed, aiming to improve the accuracy, reliability, and real-world applicability of AI in cardiology. We have considered different cardiovascular problems and approached them with AI, exploring both unsupervised and supervised ML techniques.

In Chapter 2, we explored some methodological aspects of AF risk prediction models, focusing on two main areas. The first part analyzed the discrimination and calibration performance of ML models, specifically in the context of class imbalance corrections. We showed that DL should be considered only for large sample sizes, and that undersampling the training set to improve the model's accuracy negatively affects model calibration, and that should be avoided for models developed for a clinical purpose, especially prediction. The second section of the chapter delved into the less-explored issue of determining sample sizes for DL models. This task was addressed by fitting a learning curve for CNN models trained to diagnose and predict AF. We showed that this approach can be reliably used to determine the sample size to obtain a target performance. In the third part of Chapter 2, we proposed two methodological improvements for the prediction of AF from ECG signals: the integration of ECG and tabular data and the incorporation of time-to-event and censoring information. We showed that multi-modality models have better performance compared to single-modality models.

Chapter 3 explored the use of unsupervised clustering methods for patient phenotyping and the dynamic characterization of treatment effects. The first presented study showed that clinical data, including

ECG and echocardiographic results, enabled the identification of two clusters with significant differences in genetic backgrounds and arrhythmic follow-up outcomes. The findings suggest that ECG data can effectively stratify DCM patients, potentially allowing for better risk assessment and management in clinical practice. The second study proposed a novel method for characterizing treatment responses in a survival context, using neural networks to estimate survival curves flexibly. This approach enables capturing time-dependent effects and relevant feature interactions.

Chapter 4 presented a fully automated tool for detecting the anomalous aortic origin of coronary arteries using coronary CT angiography. The tool showed high accuracy in both internal and external test sets, and has potential clinical applications, including real-time alerts for physicians and aiding in identifying high-risk cases among large patient cohorts.

Finally, Chapter 5 reported a study to asses various diagnostic modalities for detecting ATTR-CM in patients with severe aortic stenosis scheduled for TAVI. While echocardiography, CT strain, and multimodality approaches demonstrated high diagnostic accuracy, CT strain emerged as the most effective method. The study highlighted the potential of ML in leveraging routine pre-TAVI data to effectively identify concurrent ATTR-CM, providing a viable alternative to traditional methods like scintigraphy or invasive biopsies.

LIMITATIONS AND FUTURE DIRECTIONS

The primary limitation of all the studies aimed at predicting AF is the lack of validation on external datasets, which is a crucial step in developing robust prognostic models [275]. It is known that risk prediction models tend to demonstrate poorer performance when applied to new patient populations compared to their development cohorts [247]. This issue is particularly pronounced in ML models, which are prone to overfitting. The lack of generalizability is one of the main reasons why, despite the large number of ML models developed for healthcare, their implementation in clinical practice remains limited [276]. Several factors contribute to poor generalizability, including technical variations and lack of standardization in medical practice, differences in patient demographics across centers, patient genotypic and phenotypic characteristics, and hardware and software used for data acquisition [277]. While recommended, validating models on datasets beyond the original development set is often challenging due to ethical, technical, or financial barriers associated with sharing clinical data [278, 279]. One potential solution involves leveraging publicly available datasets. However, in the case of ECG data, most public datasets are designed for detection rather than prediction tasks. A promising option for predictive purposes is represented by

the UK Biobank [280], a large-scale biomedical database and research resource containing genetic, lifestyle, and health information from half a million UK participants. Although the UK Biobank follows a population-based design, which differs significantly from traditional clinical registries that focus on specific patient populations, it offers a unique opportunity to validate prediction models. Additionally, it provides a rich resource for exploring new ideas about potential predictors of future events, such as the role of genetics in ECG anomalies. Future work could focus on harnessing this resource to advance and validate the proposed prediction models.

Another limitation of the AF-prediction studies presented is the absence of explainability, a requirement that is considered essential for clinical applicability of AI systems [281, 282]. The "black-box" nature of AI models raises concerns about transparency, trust, and ethical application, as clinicians and patients often require insight into how decisions are made [283]. Explainability is critical for fostering trust among healthcare professionals, ensuring transparency in decision-making, and mitigating potential biases in AI models. For example, in our study on AAOCA detection, we integrated explainability techniques into the pipeline, demonstrating that the models were focusing on the expected part of the images. An important development would be to integrate a similar post-hoc explainability technique in the models trained for AF prediction, to make the model more interpretable. However, there is still a debate on the necessity of explainability, with some researchers highlighting that explainability is not an intrinsic necessity but rather an instrumental value, for example for model troubleshooting [284, 285]. Current explainability techniques often fail to provide sufficient clarity for patient-level decision-making and can introduce new complexities. In light of this, rigorous validation and empirical testing of AI systems might serve as more practical alternatives to achieve reliability and accountability in clinical applications.

A general limitation of the other studies reported is the small sample size of the datasets used to train and validate the models. As discussed above, access to high-quality medical data is an open challenge. The primary obstacles are privacy concerns, as sharing sensitive patient information poses significant risks, and security issues, which further complicate data exchange between institutions. Without sufficient data, models risk being undertrained, which compromises their ability to generalize effectively and perform reliably when applied to larger, more diverse populations. One potential solution to this problem is federated learning, an innovative machine learning paradigm that enables multiple institutions or entities to collaboratively train algorithms without directly sharing raw data [286]. Instead of centralizing datasets, federated learning works by keeping the data localized at its source while only sharing model updates

during the training process, minimizing privacy risks and ensuring data governance is maintained. Federated learning has shown considerable promise in the medical field [287], particularly in addressing data scarcity for tasks such as cardiovascular imaging analysis [288]. By utilizing federated learning, it is possible to access more diverse datasets, enhancing the robustness, reliability, and clinical relevance of machine learning models for cardiovascular applications [286].

As some of our results show, AI research should go in the direction of multimodality. This is for many reasons, first of all, because CVDs are heterogenous and require the integration of many sources to understand and address complex underlying mechanisms of disease and subsequent personalized treatment [31]. Indeed, this would mimic the way it is currently interpreted by doctors in clinical care, where looking at one single modality is not enough to get the complete picture of a patient. This idea is confirmed by recent research in the field, which has shifted from unimodal tools to task-agnostic, multi-modal, "foundation models" [5, 289, 290]. These models are trained on massive, diverse datasets and can be applied to numerous downstream tasks with little or no fine-tuning. These models hold immense potential to reshape healthcare and they are considered by many experts the near future of AI in healthcare.

A

SUPPLEMENTARY MATERIAL FOR CHAPTER 2

HYPERPARAMETERS TUNING

Hyperparameters are parameters that are not directly learned within estimators. We optimized hyperparameters with the RandomizedSearchCV class of the scikit-learn package. The idea is to sample a given number of candidates from a parameter space with a specified distribution. The hyperparameters with the best cross-validation AUC are then chosen. Hereafter we report the best-performing hyperparameters for XGB and LR models. Hyperparameters not reported were not tuned and used with their default value. The distributions from which we sampled candidate settings are given in the script `hyperparameters_search.py`, included in the GitHub repository of the paper.

XGB:

- `colsample_bytree`: 0.9994406226497948
- `gamma`: 4.575305537258395
- `learning_rate`: 0.02631221490906356
- `max_delta_step`: 0
- `max_depth`: 9
- `min_child_weight`: 7
- `n_estimators`: 275
- `subsample`: 0.6089510933763641

LR:

- `C`: 30.73202411565708
- `penalty`: "l2"

FEATURE IMPORTANCE

We report the regression coefficient of the LR model trained with the biggest training size (150 000 ECGs) and without RUS. Notice that 10 different models were trained with this setting since we applied 10-fold cross-validation. In Table 13 we report the coefficients of just one of the trained models. As regards XGB, we report the feature importance values directly estimated by the package (Table 14). As before,

Table 13: Standardized regression coefficients of the LR model.

Features	**Betas**
T offset	0.350
P offset	−0.318
Heart rate	0.198
QRS axis	−0.171
T axis	0.0843
QRS onset	−0.0622
P axis	0.0494
QRS offset	0.0490
P onset	−0.0351
PR interval	0.0151
QTC interval	0.0060

we consider only one of the models trained with the maximum training size and without RUS. We verified for both models that all the cross-validation models trained with the same setting shared very similar feature importance.

Table 14: Feature importance for XGB model.

Features	**Importance**
P offset	0.189
QTC interval	0.113
T axis	0.102
P onset	0.0100
P axis	0.0898
QRS axis	0.0791
T offset	0.0704
QRS onset	0.0675
PR interval	0.0660
QRS offset	0.0631
Heart rate	0.0602

B

SUPPLEMENTARY MATERIAL FOR CHAPTER 3

PCAMIX

PCAmix extends Principal Components Analysis (PCA), which is a standard multivariate analysis method for reducing the dimensions in case of a large number of variables per observation, to mixed data datasets. Similarly to PCA, the method uses a generalized singular value decomposition to obtain a geometric transformation and a subspace (by projection) that preserves the information by maximizing the variability of the projected observations. As a result, new variables called Principal Components (PCs) are obtained. PC are linear combinations of the original variables, that are ordered by maximum dispersion and by construction non-correlated among each other. By considering only the first p PCs, the dimension of the dataset is reduced. For this analysis, the first 11 PCs were considered (that is, PCs for which the proportion of explained variance/#PCs$>$ 2%).

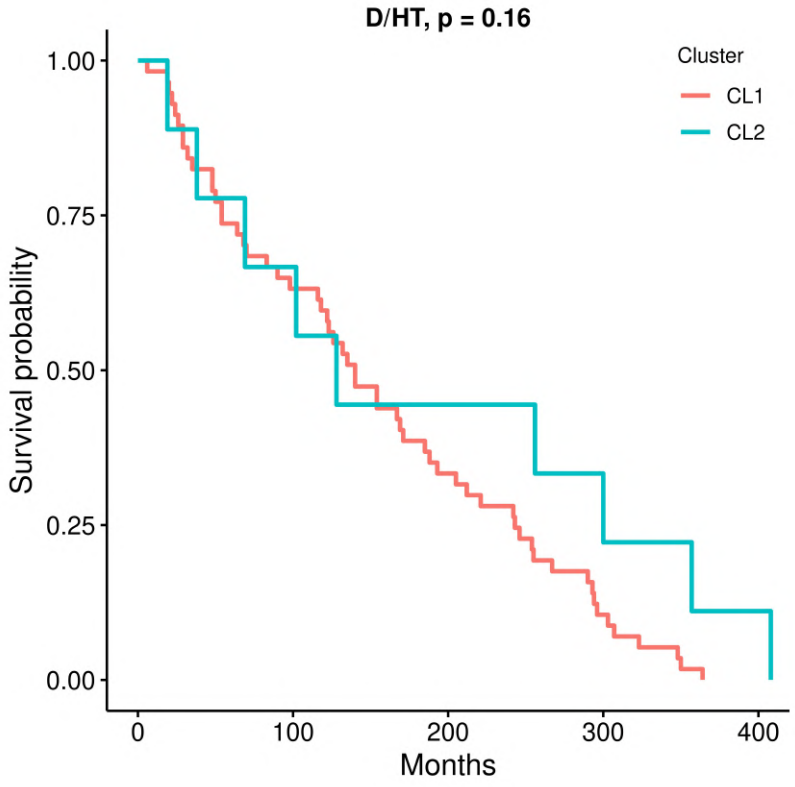


Figure 36: Kaplan-Meier curve for D/Heart Transplantation (HT) events in CL1 and CL2.

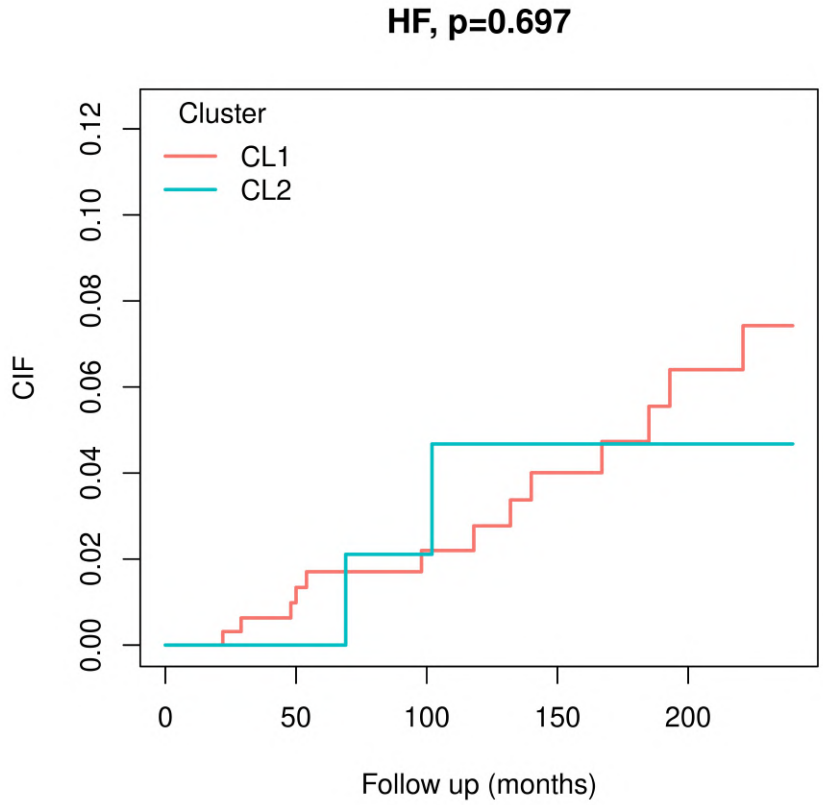


Figure 37: Cumulative incidence functions for HF events in CL1 and CL2.

Table 15: Regression coefficients of the LASSO penalized model. Example: a patient with V6 QRS interval duration 110ms, presence of true LBBB and absence of intrinsicoid deflection obtains a score of $1/(1+exp(110*0.0054+1*2.6086+0*1.1659)) = 0.0391$. Since this value is < 0.23, the subject is assigned to CL1.

Variable	**Beta**
V6 QRS interval duration	0.0054
True LBBB	2.6086
Intrinsicoid deflection	1.1659

C

SUPPLEMENTARY MATERIAL FOR CHAPTER 4

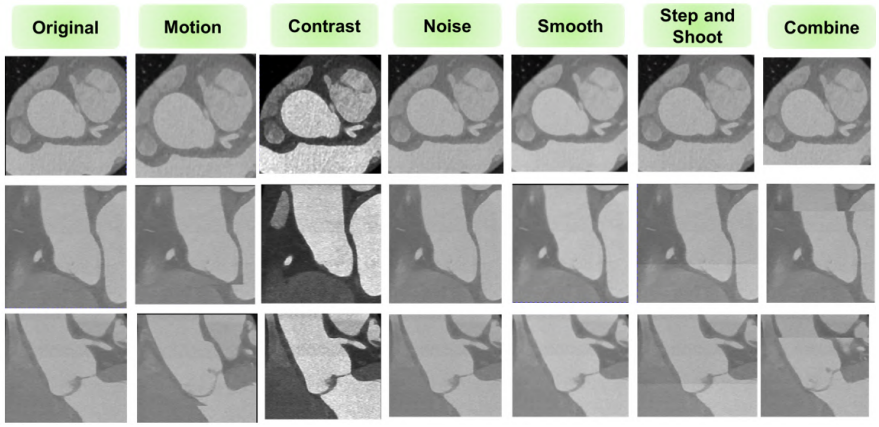


Figure 38: Original image and different augmentations applied to the image in different views.

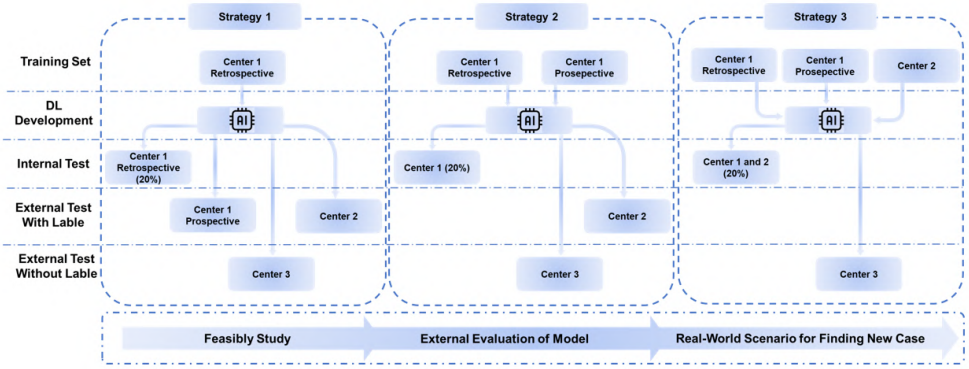


Figure 39: Possible strategies for model development. Strategy 1: described and adopted in the main study. Strategy 2: Model training is performed on the entire dataset from Bern University Hospital. The labeled dataset from Zurich University Hospital serves as an external testing dataset. The unlabeled CCTA dataset was used for external clinical evaluation, similar to Strategy 1. Strategy 3: Model training is performed on all datasets with labels, including data from Bern and Zurich University Hospitals. External model performance is evaluated in the unlabeled dataset.

Table 16: Summary statistics of the number of patients and images in each dataset for different classification tasks.

		Anomaly Detection		
	# All patients	# All images	#AAOCA patients	#AAOCA images
Internal Test Retrospective cohort	536	1567	147	598
Internal Test Prospective cohort	359	1066	58	319
External Test	483	497	130	139
External Clinical Evaluation	998	998	Unknown	Unknown
		Origin Classification		
	# All patients	# All images	#L-AAOCA patients	# L-AAOCA images
Internal Test Retrospective cohort	145	585	54	207
Internal Test Prospective cohort	57	309	15	48
External Test	126	135	64	68
		Risk Classification		
	# All patients	# All images	#High-risk patients	# High risk- images
Internal Test Retrospective cohort	144	582	107	465
Internal Test Prospective cohort	57	309	45	275
External Test	126	135	74	80

Table 17: Performance metrics of individual models, mean metrics across all models, and ensemble metrics obtained by combining the predictions of individual models for the anomaly detection task, evaluated on internal and external testing datasets.

		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Ensemble
Test internal	**ROC AUC**	0.993	0.990	0.994	0.994	0.995	0.993	0.998
	Sensitivity	0.978	0.925	0.984	0.956	0.994	0.967	0.987
	Specificity	0.984	0.992	0.979	0.987	0.976	0.983	0.989
	F1-score	0.970	0.952	0.968	0.962	0.969	0.964	0.981
	PPV	0.963	0.980	0.952	0.968	0.946	0.962	0.975
	AUPR	0.971	0.984	0.982	0.982	0.978	0.979	0.996
	Accuracy	0.982	0.972	0.980	0.977	0.981	0.979	0.989
Test external	**ROC AUC**	0.994	0.997	0.999	0.999	0.998	0.997	0.999
	Sensitivity	0.928	0.942	0.950	0.964	0.971	0.951	0.957
	Specificity	0.994	0.994	0.994	0.997	0.980	0.992	1.
	F1-score	0.956	0.963	0.967	0.978	0.961	0.965	0.978
	PPV	0.985	0.985	0.985	0.993	0.951	0.980	1.
	AUPR	0.992	0.994	0.997	0.997	0.994	0.995	0.999
	Accuracy	0.976	0.980	0.982	0.988	0.978	0.981	0.988

Table 18: Performance metrics of individual models, mean metrics across all models, and ensemble metrics obtained by combining the predictions of individual models for the origin classification task, evaluated on internal and external testing datasets.

		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Ensemble
Test internal	**ROC AUC**	1.000	0.994	0.997	0.999	0.999	0.998	0.999
	Sensitivity	0.938	0.875	0.958	0.938	0.938	0.929	0.938
	Specificity	1.000	0.989	0.996	1.000	1.000	0.997	1.000
	F1-score	0.968	0.903	0.968	0.968	0.968	0.955	0.968
	PPV	1.000	0.933	0.979	1.000	1.000	0.982	1.000
	AUPR	0.999	0.976	0.988	0.996	0.995	0.991	0.997
	Accuracy	0.990	0.971	0.990	0.990	0.990	0.986	0.990
Test external	**ROC AUC**	0.990	0.988	0.997	0.978	0.997	0.990	0.999
	Sensitivity	0.970	0.955	0.940	0.940	0.985	0.958	0.955
	Specificity	1.000	0.985	1.000	0.970	0.985	0.988	1.000
	F1-score	0.985	0.970	0.969	0.955	0.985	0.973	0.977
	PPV	1.000	0.985	1.000	0.969	0.985	0.988	1.000
	AUPR	0.994	0.992	0.997	0.984	0.997	0.993	0.999
	Accuracy	0.985	0.970	0.970	0.955	0.985	0.973	0.978

Table 19: Performance metrics of individual models, mean metrics across all models, and ensemble metrics obtained by combining the predictions of individual models for the risk classification task, evaluated on internal and external testing datasets.

		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Ensemble
Test internal	**ROC AUC**	0.994	0.984	1.000	1.000	1.000	0.996	0.999
	Sensitivity	0.982	0.902	0.996	0.967	0.978	0.965	0.989
	Specificity	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	F1-score	0.991	0.948	0.998	0.983	0.989	0.982	0.995
	PPV	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	AUPR	0.999	0.998	1.000	1.000	1.000	0.999	1.000
	Accuracy	0.984	0.913	0.997	0.971	0.981	0.969	0.990
Test external	**ROC AUC**	0.989	0.980	0.999	0.998	0.997	0.993	0.996
	Sensitivity	0.975	0.875	0.975	0.950	0.975	0.950	0.962
	Specificity	0.963	0.944	0.981	0.963	0.963	0.963	0.963
	F1-score	0.975	0.915	0.981	0.962	0.975	0.962	0.969
	PPV	0.975	0.959	0.987	0.974	0.975	0.974	0.975
	AUPR	0.992	0.985	0.999	0.999	0.998	0.995	0.997
	Accuracy	0.970	0.903	0.978	0.955	0.970	0.955	0.963

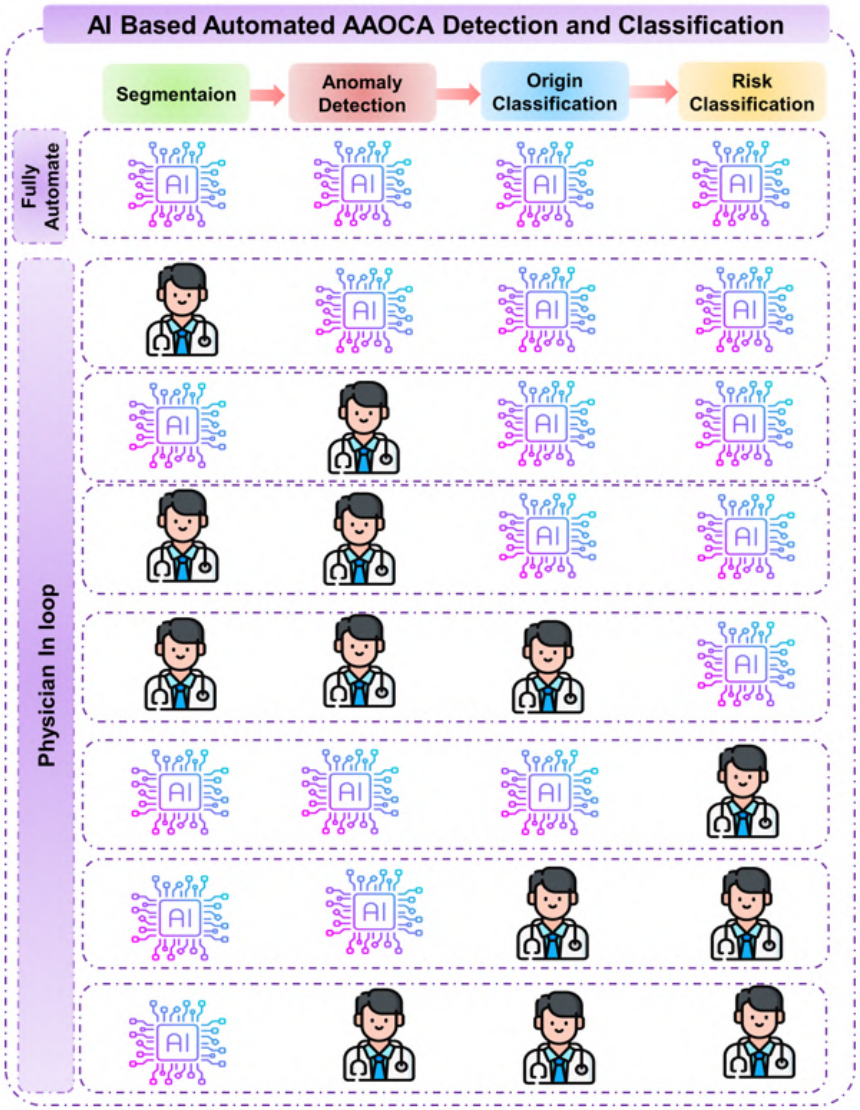


Figure 40: Different possibilities for using the developed AI model in clinical scenarios include fully automated applications and physician-in-the-loop systems.

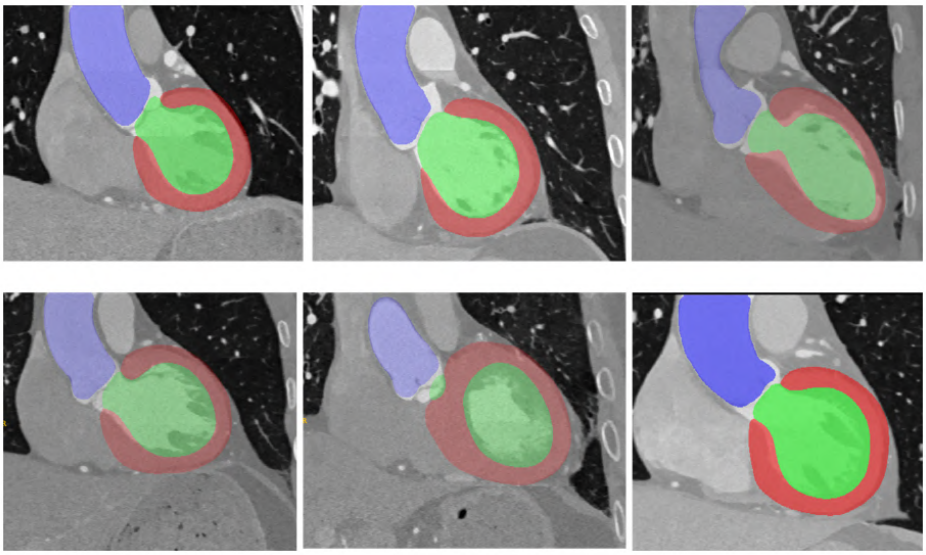


Figure 41: Six different cases with their corresponding segmentations of the aorta and left ventricle.

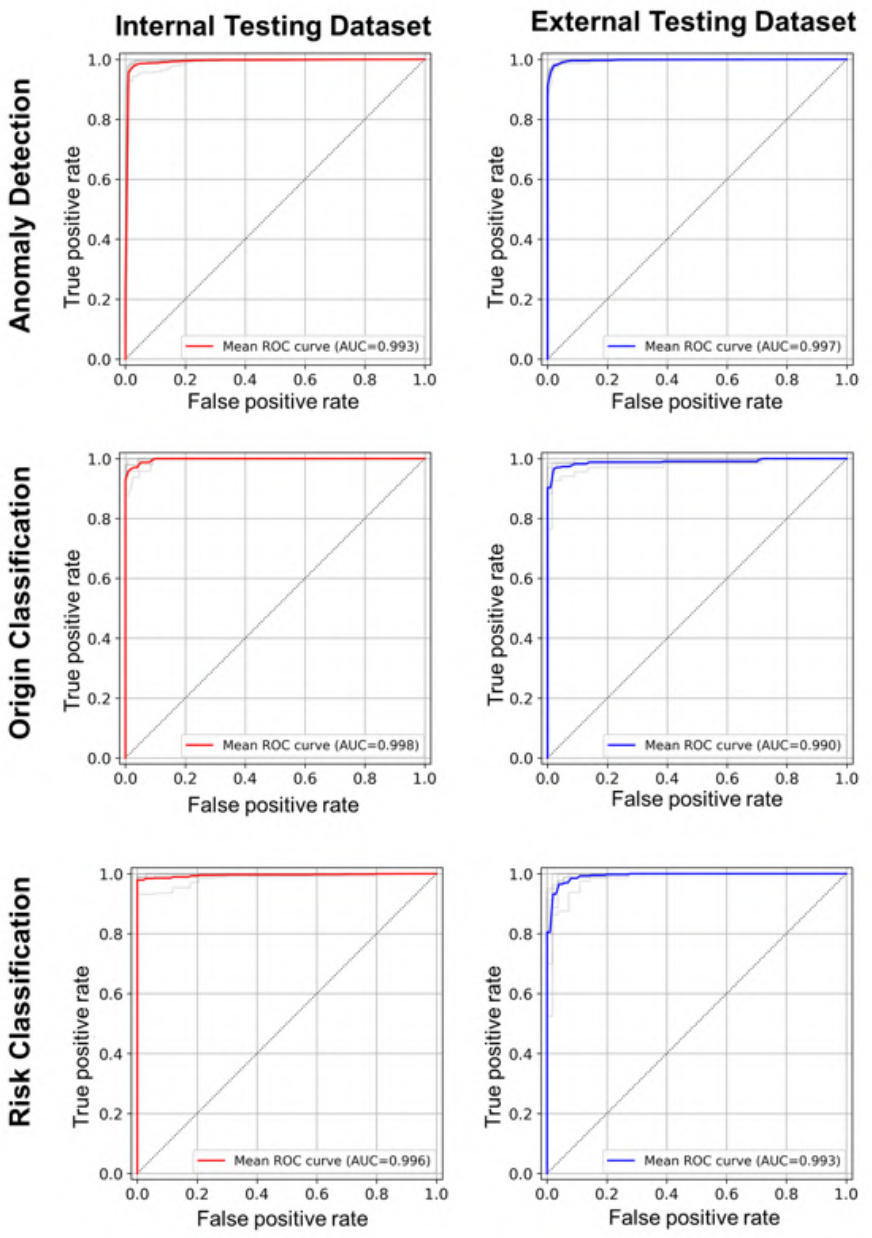


Figure 42: Mean ROC curves of 5 different folds across different Tasks for various datasets.

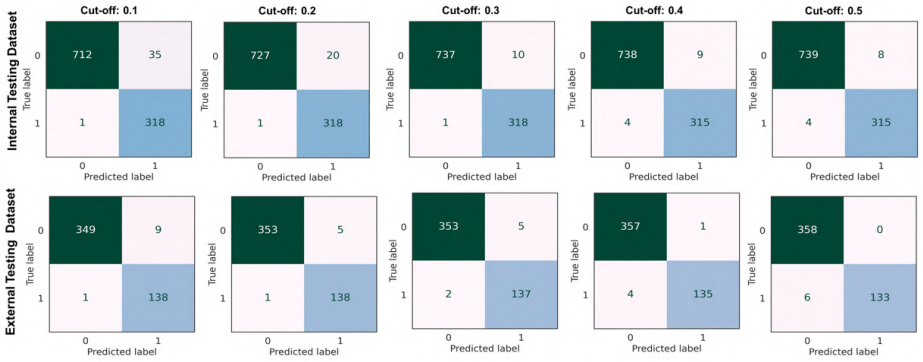


Figure 43: Confusion matrices of the ensemble model in Anomaly Detection for various datasets at different cut-off points.

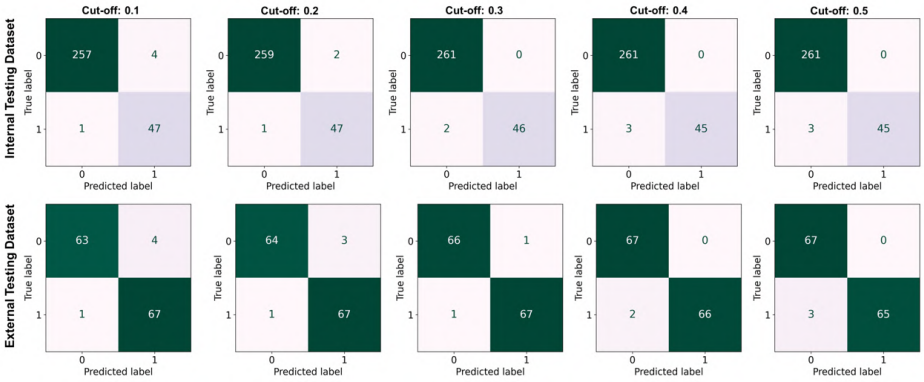


Figure 44: Confusion matrices of the ensemble model in Origin Classification for various datasets at different cut-off points.

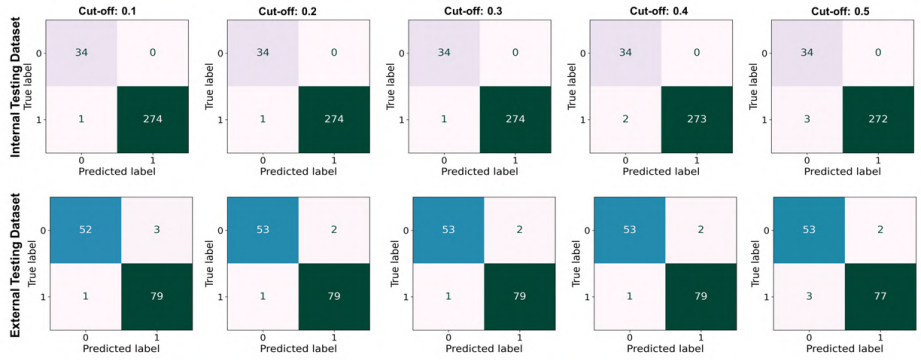


Figure 45: Confusion matrices of the ensemble model in Risk Classification for various datasets at different cut-off points.

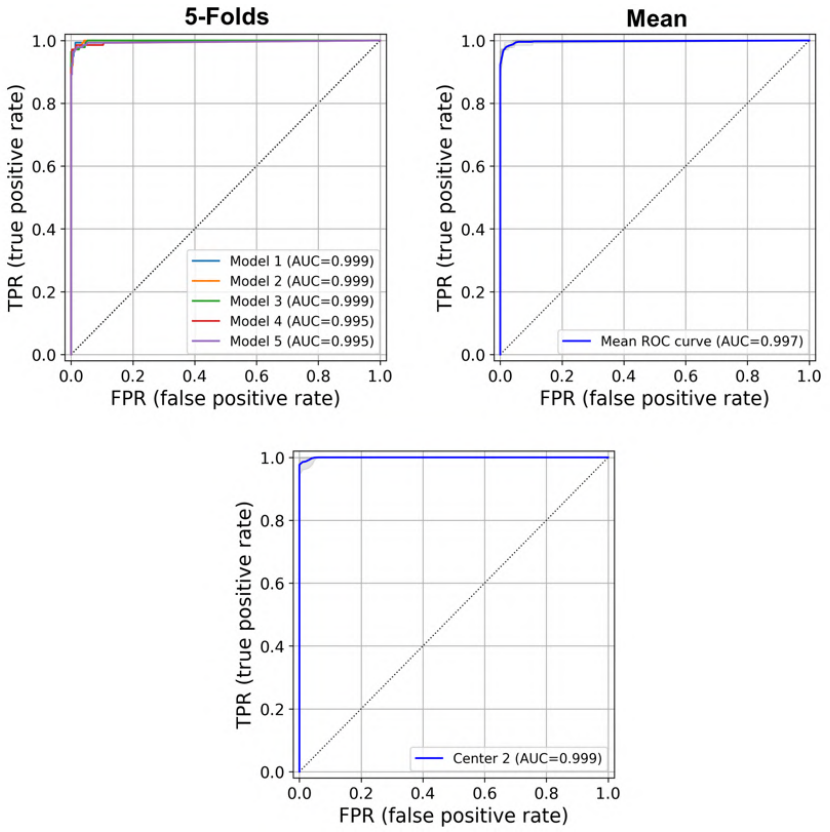


Figure 46: ROC curves of 5 folds, mean, and an ensemble of strategy 2 in the external test dataset from center 2 (Anomaly Detection task).

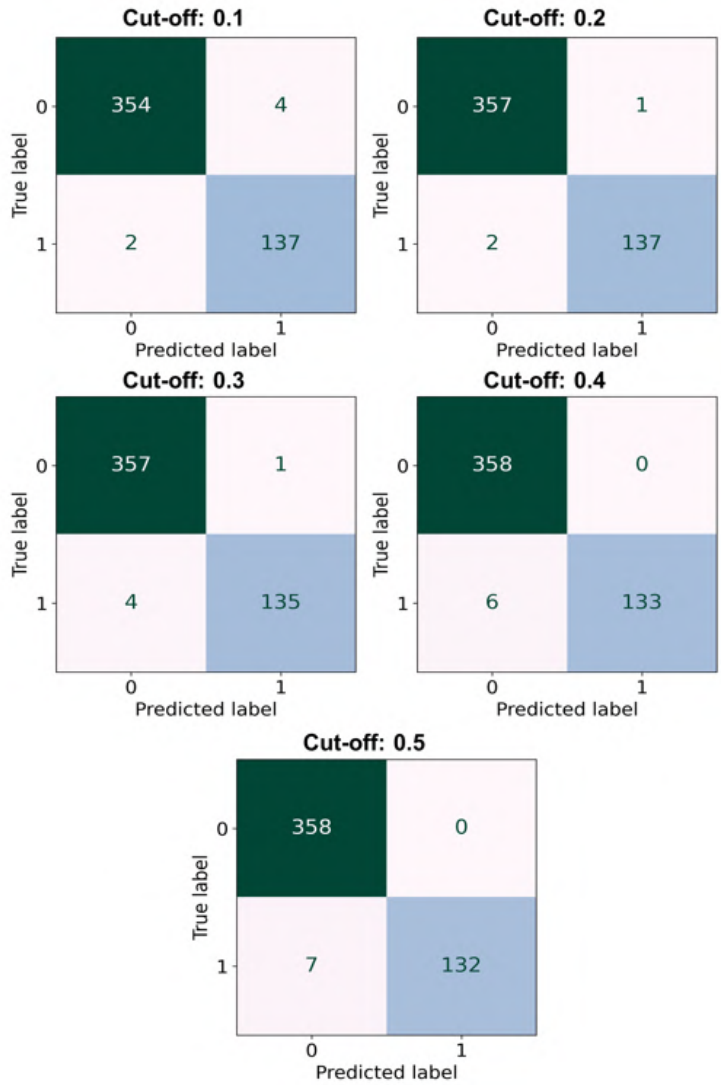


Figure 47: Confusion matrices of ensemble model in Anomaly Detection for strategy 2 at different cut-off points.

BIBLIOGRAPHY

[1] Muthiah Vaduganathan, George A. Mensah, Justine Varieur Turco, Valentin Fuster, and Gregory A. Roth. "The Global Burden of Cardiovascular Diseases and Risk." In: *Journal of the American College of Cardiology* 80.25 (Dec. 2022), pp. 2361–2371. DOI: 10.1016/j.jacc.2022.11.005.

[2] Gregory A. Roth et al. "Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019." In: *Journal of the American College of Cardiology* 76.25 (Dec. 2020), pp. 2982–3021. DOI: 10.1016/j.jacc.2020.11.010.

[3] Dunbar et al. "Projected Costs of Informal Caregiving for Cardiovascular Disease: 2015 to 2035: A Policy Statement From the American Heart Association | Circulation." In: *Circulation* 137.19 (May 2018), e558–e577. DOI: https://doi.org/10.1161/CIR.0000000000000570.

[4] Ramon Luengo-Fernandez et al. "Economic burden of cardiovascular diseases in the European Union: a population-based cost study." In: *European Heart Journal* 44.45 (Dec. 2023), pp. 4752–4767. DOI: 10.1093/eurheartj/ehad583.

[5] Michael D. Howell, Greg S. Corrado, and Karen B. DeSalvo. "Three Epochs of Artificial Intelligence in Health Care." In: *JAMA* 331.3 (Jan. 2024), pp. 242–244. DOI: 10.1001/jama.2023.25057.

[6] U.S. Food and Drug Administration (FDA). *Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices*. Aug. 2024. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices.

[7] M. D. Pierre Elias et al. "Artificial Intelligence for Cardiovascular Care - Part 1: Advances: JACC Review Topic of the Week." In: *Journal of the American College of Cardiology* (Mar. 2024). DOI: 10.1016/j.jacc.2024.03.400.

[8] Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network." In: *Nature Medicine* 25.1 (Jan. 2019), pp. 65–69. DOI: 10.1038/s41591-018-0268-3.

[9] Antônio H. Ribeiro et al. "Automatic diagnosis of the 12-lead ECG using a deep neural network." In: *Nature Communications* 11.1 (Apr. 2020), p. 1760. DOI: 10.1038/s41467-020-15432-4.

[10] Hongling Zhu et al. "Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study." In: *The Lancet Digital Health* 2.7 (July 2020), e348–e357. DOI: 10.1016/S2589-7500(20)30107-2.

[11] Zachi I. Attia et al. "An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction." In: *The Lancet* 394.10201 (Sept. 2019), pp. 861–867. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(19)31721-0.

[12] Sushravya Raghunath et al. "Deep Neural Networks Can Predict New-Onset Atrial Fibrillation From the 12-Lead ECG and Help Identify Those at Risk of Atrial Fibrillation–Related Stroke." In: *Circulation* 143.13 (Mar. 2021), pp. 1287–1298. DOI: 10.1161/CIRCULATIONAHA.120.047829.

[13] Zachi I. Attia et al. "Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram." In: *Nature Medicine* 25.1 (Jan. 2019), pp. 70–74. DOI: 10.1038/s41591-018-0240-2.

[14] Veer Sangha et al. "Detection of Left Ventricular Systolic Dysfunction From Electrocardiographic Images." In: *Circulation* 148.9 (Aug. 2023), pp. 765–777. DOI: 10.1161/CIRCULATIONAHA.122.062646.

[15] Michal Cohen-Shelly et al. "Electrocardiogram screening for aortic valve stenosis using artificial intelligence." In: *European Heart Journal* 42.30 (Aug. 2021), pp. 2885–2896. DOI: 10.1093/eurheartj/ehab153.

[16] J. Martijn Bos, Zachi I. Attia, David E. Albert, Peter A. Noseworthy, Paul A. Friedman, and Michael J. Ackerman. "Use of Artificial Intelligence and Deep Neural Networks in Evaluation of Patients With Electrocardiographically Concealed Long QT Syndrome From the Surface 12-Lead Electrocardiogram." In: *JAMA Cardiology* 6.5 (May 2021), pp. 532–538. DOI: 10.1001/jamacardio.2020.7422.

[17] Conner D. Galloway et al. "Development and Validation of a Deep-Learning Model to Screen for Hyperkalemia From the Electrocardiogram." In: *JAMA Cardiology* 4.5 (May 2019), pp. 428–436. DOI: 10.1001/jamacardio.2019.0640.

[18] Joon-myoung Kwon, Younghoon Cho, Ki-Hyun Jeon, Soohyun Cho, Kyung-Hee Kim, Seung Don Baek, Soomin Jeung, Jinsik Park, and Byung-Hee Oh. "A deep learning algorithm to detect anaemia with ECGs: a retrospective, multicentre study." In: *The Lancet Digital Health* 2.7 (July 2020), e358–e367. DOI: 10.1016/S2589-7500(20)30108-4.

[19] Akhil Narang et al. "Utility of a Deep-Learning Algorithm to Guide Novices to Acquire Echocardiograms for Limited Diagnostic Use." In: *JAMA Cardiology* 6.6 (June 2021), pp. 624–632. DOI: 10.1001/jamacardio.2021.0185.

[20] Neal Yuan, Ishan Jain, Neeraj Rattehalli, Bryan He, Charles Pollick, David Liang, Paul Heidenreich, James Zou, Susan Cheng, and David Ouyang. "Systematic Quantification of Sources of Variation in Ejection Fraction Calculation Using Deep Learning." In: *JACC: Cardiovascular Imaging* 14.11 (Nov. 2021), pp. 2260–2262. DOI: 10.1016/j.jcmg.2021.06.018.

[21] Nikki van der Velde, H. Carlijne Hassing, Brendan J. Bakker, Piotr A. Wielopolski, R. Marc Lebel, Martin A. Janich, Isabella Kardys, Ricardo P. J. Budde, and Alexander Hirsch. "Improvement of late gadolinium enhancement image quality using a deep learning–based reconstruction algorithm and its influence on myocardial scar quantification." In: *European Radiology* 31.6 (June 2021), pp. 3846–3855. DOI: 10.1007/s00330-020-07461-w.

[22] Wenjia Bai et al. "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks." In: *Journal of Cardiovascular Magnetic Resonance* 20.1 (Sept. 2018), p. 65. DOI: 10.1186/s12968-018-0471-x.

[23] Sanne G. M. van Velzen et al. "Deep Learning for Automatic Calcium Scoring in CT: Validation Using Multiple Cardiac CT and Chest CT Protocols." In: *Radiology* 295.1 (Apr. 2020), pp. 66–79. DOI: 10.1148/radiol.2020191621.

[24] Andrew Lin et al. "Deep learning-enabled coronary CT angiography for plaque and stenosis quantification and cardiac risk prediction: an international multicentre study." In: *The Lancet Digital Health* 4.4 (Apr. 2022), e256–e265. DOI: 10.1016/S2589-7500(22)00022-X.

[25] Allison W. Peng, Ramzi Dudum, Sneha S. Jain, David J. Maron, Bhavik N. Patel, Nishith Khandwala, David Eng, Akshay S. Chaudhari, Alexander T. Sandhu, and Fatima Rodriguez. "Association of Coronary Artery Calcium Detected by Routine Ungated CT Imaging With Cardiovascular Outcomes." In: *Journal of the American College of Cardiology* 82.12 (Sept. 2023), pp. 1192–1202. DOI: 10.1016/j.jacc.2023.06.040.

[26] Ghalib A. Bello et al. "Deep-learning cardiac motion analysis for human survival prediction." In: *Nature Machine Intelligence* 1.2 (Feb. 2019), pp. 95–104. DOI: 10.1038/s42256-019-0019-2.

[27] Maximilian J. Bauer, Nejva Nano, Rafael Adolf, Albrecht Will, Eva Hendrich, Stefan A. Martinoff, and Martin Hadamitzky. "Prognostic Value of Machine Learning–based Time-to-Event

Analysis Using Coronary CT Angiography in Patients with Suspected Coronary Artery Disease." In: *Radiology: Cardiothoracic Imaging* 5.2 (Apr. 2023). Publisher: Radiological Society of North America, e220107. DOI: 10.1148/ryct.220107.

[28] Kun-Hsing Yu, Andrew L. Beam, and Isaac S. Kohane. "Artificial intelligence in healthcare." In: *Nature Biomedical Engineering* 2.10 (Oct. 2018), pp. 719–731. DOI: 10.1038/s41551-018-0305-z.

[29] Sammy Chouffani El Fassi, Adonis Abdullah, Ying Fang, Sarabesh Natarajan, Awab Bin Masroor, Naya Kayali, Simran Prakash, and Gail E. Henderson. "Not all AI health tools with regulatory authorization are clinically validated." In: *Nature Medicine* (Aug. 2024). DOI: 10.1038/s41591-024-03203-3.

[30] Partho P. Sengupta et al. "Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation (PRIME): A Checklist: Reviewed by the American College of Cardiology Healthcare Innovation Council." In: *JACC: Cardiovascular Imaging* 13.9 (Sept. 2020), pp. 2017–2035. DOI: 10.1016/j.jcmg.2020.07.015.

[31] Sneha Jain et al. "Artificial Intelligence in Cardiovascular Care — Part 2: Applications: JACC Review Topic of the Week." In: *Journal of the American College of Cardiology* (Apr. 2024). ISSN: 0735-1097. DOI: 10.1016/j.jacc.2024.03.401.

[32] Stuart J. Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Third edition, Global edition. Prentice Hall series in artificial intelligence. Pearson, 2016. ISBN: 978-1-292-15396-4.

[33] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. Adaptive computation and machine learning. Cambridge, Mass: The MIT press, 2016. ISBN: 978-0-262-03561-3.

[34] Zhi-Hua Zhou. *Machine learning*. Singapore: Springer, 2021. ISBN: 9789811519673.

[35] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. Adaptive computation and machine learning. Cambridge, Mass. London: The MIT Press, 2012. ISBN: 978-0-262-01825-8.

[36] Asma Chebli, Akila Djebbar, and Hayet Farida Marouani. "Semi-Supervised Learning for Medical Application: A Survey." In: *2018 International Conference on Applied Smart Systems (ICASS)*. Nov. 2018, pp. 1–9. DOI: 10.1109/ICASS.2018.8651980.

[37] Samuel Budd, Emma C. Robinson, and Bernhard Kainz. "A survey on active learning and human-in-the-loop deep learning for medical image analysis." In: *Medical Image Analysis* 71 (July 2021), p. 102062. ISSN: 1361-8415. DOI: 10.1016/j.media.2021.102062.

[38] Rayan Krishnan, Pranav Rajpurkar, and Eric J. Topol. "Self-supervised learning in medicine and healthcare." In: *Nature Biomedical Engineering* 6.12 (Dec. 2022), pp. 1346–1352. ISSN: 2157-846X. DOI: 10.1038/s41551-022-00914-1.

[39] S. B. Kotsiantis. "Decision trees: a recent overview." In: *Artificial Intelligence Review* 39.4 (Apr. 2013), pp. 261–283. ISSN: 1573-7462. DOI: 10.1007/s10462-011-9272-4.

[40] Corinna Cortes and Vladimir Vapnik. "Support-vector networks." In: *Machine Learning* 20.3 (Sept. 1995), pp. 273–297. ISSN: 1573-0565. DOI: 10.1007/BF00994018.

[41] Anders Krogh. "What are artificial neural networks?" In: *Nature Biotechnology* 26.2 (Feb. 2008), pp. 195–197. ISSN: 1546-1696. DOI: 10.1038/nbt1386.

[42] T. Cover and P. Hart. "Nearest neighbor pattern classification." In: *IEEE Transactions on Information Theory* 13.1 (Jan. 1967), pp. 21–27. ISSN: 1557-9654. DOI: 10.1109/TIT.1967.1053964.

[43] Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. "A survey on ensemble learning." In: *Frontiers of Computer Science* 14.2 (Apr. 2020), pp. 241–258. ISSN: 2095-2236. DOI: 10.1007/s11704-019-8208-z.

[44] Alexey Natekin and Alois Knoll. "Gradient boosting machines, a tutorial." In: *Frontiers in Neurorobotics* 7 (Dec. 2013). ISSN: 1662-5218. DOI: 10.3389/fnbot.2013.00021.

[45] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[46] Leo Breiman. "Bagging predictors." In: *Machine Learning* 24.2 (Aug. 1996), pp. 123–140. ISSN: 1573-0565. DOI: 10.1007/BF00058655.

[47] Leo Breiman. "Random Forests." In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324.

[48] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." In: *Nature* 521.7553 (May 2015), pp. 436–444. ISSN: 1476-4687. DOI: 10.1038/nature14539.

[49] Warren S. McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity." In: *The bulletin of mathematical biophysics* 5.4 (Dec. 1943), pp. 115–133. ISSN: 1522-9602. DOI: 10.1007/BF02478259.

[50] Christopher M. Bishop and Hugh Bishop. *Deep Learning*. Springer, 2024. ISBN: 978-3-031-45468-4.

[51] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. "Deep Learning for Computer Vision: A Brief Review." In: *Computational Intelligence and Neuroscience* 2018.1 (2018), p. 7068349. ISSN: 1687-5273. DOI: 10.1155/2018/7068349.

[52] Dominik Scherer, Andreas Müller, and Sven Behnke. "Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition." In: *Artificial Neural Networks – ICANN 2010*. Ed. by Konstantinos Diamantaras, Wlodek Duch, and Lazaros S. Iliadis. Berlin, Heidelberg: Springer, 2010. DOI: 10.1007/978-3-642-15825-4_10.

[53] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." In: *Proceedings of the IEEE* 86.11 (Nov. 1998), pp. 2278–2324. ISSN: 1558-2256. DOI: 10.1109/5.726791.

[54] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. Jan. 2017. DOI: 10.48550/arXiv.1412.6980.

[55] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." In: *Nature* 323.6088 (Oct. 1986), pp. 533–536. ISSN: 1476-4687. DOI: 10.1038/323533a0.

[56] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." In: *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958. ISSN: 1533-7928.

[57] Lutz Prechelt. "Early Stopping — But When?" In: *Neural Networks: Tricks of the Trade: Second Edition*. Ed. by Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller. Berlin, Heidelberg: Springer, 2012, pp. 53–67. ISBN: 978-3-642-35289-8. DOI: 10.1007/978-3-642-35289-8_5.

[58] Jelena Kornej, Christin S. Börschel, Emelia J. Benjamin, and Renate B. Schnabel. "Epidemiology of Atrial Fibrillation in the 21st Century." In: *Circulation Research* 127.1 (June 2020), pp. 4–20. DOI: 10.1161/CIRCRESAHA.120.316340.

[59] Y. Miyasaka, M.E. Barnes, B.J. Gersh, S.S. Cha, K.R. Bailey, W.P. Abhayaratna, J.B. Seward, and T.S.M. Tsang. "Secular trends in incidence of atrial fibrillation in Olmsted County, Minnesota, 1980 to 2000, and implications on the projections for future prevalence." In: *Circulation* 114.2 (2006), pp. 119–125. DOI: 10.1161/CIRCULATIONAHA.105.595140.

[60] Bouwe P. Krijthe, Anton Kunst, Emelia J. Benjamin, Gregory Y.H. Lip, Oscar H. Franco, Albert Hofman, Jacqueline C.M. Witteman, Bruno H. Stricker, and Jan Heeringa. "Projections on the number of individuals with atrial fibrillation in the European Union, from 2000 to 2060." In: *European Heart Journal* 34.35 (Sept. 2013), pp. 2746–2751. ISSN: 0195-668X. DOI: 10.1093/eurheartj/eht280.

[61] Faisal Rahman, Gene F. Kwan, and Emelia J. Benjamin. "Global epidemiology of atrial fibrillation." In: *Nature Reviews Cardiology* 11.11 (Nov. 2014), pp. 639–654. ISSN: 1759-5010. DOI: 10.1038/nrcardio.2014.118.

[62] Polychronis E. Dilaveris and Harold L. Kennedy. "Silent atrial fibrillation: epidemiology, diagnosis, and clinical impact." In: *Clinical Cardiology* 40.6 (2017), pp. 413–418. ISSN: 1932-8737. DOI: 10.1002/clc.22667.

[63] Paulus Kirchhof et al. "2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS." In: *European Heart Journal* 37.38 (Oct. 2016), pp. 2893–2962. ISSN: 0195-668X. DOI: 10.1093/eurheartj/ehw210.

[64] Henrik Stig Jørgensen, Hirofumi Nakayama, Jakob Reith, Hans Otto Raaschou, and Tom Skyhøj Olsen. "Acute Stroke With Atrial Fibrillation." In: *Stroke* 27.10 (Oct. 1996), pp. 1765–1769. DOI: 10.1161/01.STR.27.10.1765.

[65] Emer R. McGrath, Moira K. Kapral, Jiming Fang, John W. Eikelboom, Aengus ó Conghaile, Michelle Canavan, Martin J. O'Donnell, and Investigators of the Ontario Stroke Registry. "Association of atrial fibrillation with mortality and disability after ischemic stroke." In: *Neurology* 81.9 (Aug. 2013), pp. 825–832. ISSN: 1526-632X. DOI: 10.1212/WNL.0b013e3182a2cc15.

[66] Joy Liao, Zahira Khalid, Ciaran Scallan, Carlos Morillo, and Martin O'Donnell. "Noninvasive Cardiac Monitoring for Detecting Paroxysmal Atrial Fibrillation or Flutter After Acute Ischemic Stroke." In: *Stroke* 38.11 (Nov. 2007), pp. 2935–2940. DOI: 10.1161/STROKEAHA.106.478685.

[67] Walter P. Wodchis, R. Sacha Bhatia, Kori Leblanc, Nazanin Meshkat, and Dante Morra. "A Review of the Cost of Atrial Fibrillation." In: *Value in Health* 15.2 (Mar. 2012), pp. 240–248. ISSN: 1098-3015. DOI: 10.1016/j.jval.2011.09.009.

[68] Massimo Salvi, Madhav R. Acharya, Silvia Seoni, Oliver Faust, Ru-San Tan, Prabal Datta Barua, Salvador García, Filippo Molinari, and U. Rajendra Acharya. "Artificial intelligence for atrial fibrillation detection, prediction, and treatment: A systematic review of the last decade (2013–2023)." In: *WIREs Data Mining and Knowledge Discovery* 14.3 (2024), e1530. ISSN: 1942-4795. DOI: 10.1002/widm.1530.

[69] Oliver Faust, Edward J. Ciaccio, and U. Rajendra Acharya. "A Review of Atrial Fibrillation Detection Methods as a Service." In: *International Journal of Environmental Research and Public Health* 17.9 (Jan. 2020), p. 3093. ISSN: 1660-4601. DOI: 10.3390/ijerph17093093.

[70] Vias Markides and Richard J. Schilling. "Atrial fibrillation: classification, pathophysiology, mechanisms and drug treatment." In: *Heart* 89.8 (Aug. 2003), pp. 939–943. ISSN: 1355-6037, 1468-201X. DOI: 10.1136/heart.89.8.939.

[71] Heinz-Peter Schultheiss et al. "Dilated cardiomyopathy." In: *Nature Reviews Disease Primers* 5.1 (May 2019), pp. 1–19. DOI: 10.1038/s41572-019-0084-1.

[72] Tiziana Ciarambino, Giovanni Menna, Gennaro Sansone, and Mauro Giordano. "Cardiomyopathies: An Overview." In: *International Journal of Molecular Sciences* 22.14 (Jan. 2021), p. 7722. ISSN: 1422-0067. DOI: 10.3390/ijms22147722.

[73] Perry Elliott et al. "Classification of the cardiomyopathies: a position statement from the european society of cardiology working group on myocardial and pericardial diseases." In: *European Heart Journal* 29.2 (Jan. 2008), pp. 270–276. DOI: 10.1093/eurheartj/ehm342.

[74] Robert G. Weintraub, Christopher Semsarian, and Peter Macdonald. "Dilated cardiomyopathy." In: *The Lancet* 390.10092 (July 2017), pp. 400–414. DOI: 10.1016/S0140-6736(16)31713-5.

[75] G. William Dec and Valentin Fuster. "Idiopathic Dilated Cardiomyopathy." In: *New England Journal of Medicine* 331.23 (Dec. 1994), pp. 1564–1575. ISSN: 0028-4793. DOI: 10.1056/NEJM199412083312307.

[76] Mario Gaudino et al. "Management of Adults With Anomalous Aortic Origin of the Coronary Arteries." In: *Journal of the American College of Cardiology* 82.21 (Nov. 2023), pp. 2034–2053. ISSN: 07351097. DOI: 10.1016/j.jacc.2023.08.012.

[77] Barry J. Maron, Joseph J. Doerer, Tammy S. Haas, David M. Tierney, and Frederick O. Mueller. "Sudden Deaths in Young Competitive Athletes." In: *Circulation* 119.8 (Mar. 2009), pp. 1085–1092. DOI: 10.1161/CIRCULATIONAHA.108.804617.

[78] Robert E. Eckart, Stephanie L. Scoville, Charles L. Campbell, Eric A. Shry, Karl C. Stajduhar, Robert N. Potter, Lisa A. Pearse, and Renu Virmani. "Sudden Death in Young Adults: A 25-Year Review of Autopsies in Military Recruits." In: *Annals of Internal Medicine* 141.11 (Dec. 2004), pp. 829–834. ISSN: 0003-4819. DOI: 10.7326/0003-4819-141-11-200412070-00005.

[79] Silvana Molossi, Luis E. Martínez-Bravo, and Carlos M. Mery. "Anomalous Aortic Origin of a Coronary Artery." In: *Methodist DeBakey Cardiovascular J* 15.2 (Apr. 2019). ISSN: 1947-6094. DOI: 10.14797/mdcj-15-2-111.

[80] Christoph Gräni et al. "First report from the European registry for anomalous aortic origin of coronary artery (EURO-AAOCA)." In: *Interdisciplinary CardioVascular and Thoracic Surgery* 38.5 (May 2024), ivae074. ISSN: 2753-670X. DOI: 10.1093/icvts/ivae074.

[81] Helmut Baumgartner et al. "2020 ESC Guidelines for the management of adult congenital heart disease: The Task Force for the management of adult congenital heart disease of the European Society of Cardiology (ESC). Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC), International Society for Adult Congenital Heart Disease (ISACHD)." In: *European Heart Journal* 42.6 (Feb. 2021), pp. 563–645. ISSN: 0195-668X. DOI: 10.1093/eurheartj/ehaa554.

[82] Louhai Alwan et al. "Current and Evolving Multimodality Cardiac Imaging in Managing Transthyretin Amyloid Cardiomyopathy." In: *JACC: Cardiovascular Imaging* 17.2 (Feb. 2024), pp. 195–211. DOI: 10.1016/j.jcmg.2023.10.010.

[83] Milagros Pereyra Pietri et al. "The prognostic value of artificial intelligence to predict cardiac amyloidosis in patients with severe aortic stenosis undergoing transcatheter aortic valve replacement." In: *European Heart Journal - Digital Health* 5.3 (May 2024), pp. 295–302. DOI: 10.1093/ehjdh/ztae022.

[84] Gioele Fabbri, Matteo Serenelli, Anna Cantone, Federico Sanguettoli, and Claudio Rapezzi. "Transthyretin amyloidosis in aortic stenosis: clinical and therapeutic implications." In: *European Heart Journal Supplements* 23.Supplement_E (Oct. 2021), E128–E132. DOI: 10.1093/eurheartj/suab107.

[85] Adam Castaño et al. "Unveiling transthyretin cardiac amyloidosis and its predictors among elderly patients with severe aortic stenosis undergoing transcatheter aortic valve replacement." In: *European Heart Journal* 38.38 (Oct. 2017), pp. 2879–2887. DOI: 10.1093/eurheartj/ehx350.

[86] Christian Nitsche et al. "Prevalence and Outcomes of Concomitant Aortic Stenosis and Cardiac Amyloidosis." In: *Journal of the American College of Cardiology* 77.2 (Jan. 2021), pp. 128–139. DOI: 10.1016/j.jacc.2020.11.006.

[87] Julien Ternacle et al. "Aortic Stenosis and Cardiac Amyloidosis." In: *Journal of the American College of Cardiology* 74.21 (Nov. 2019), pp. 2638–2651. DOI: 10.1016/j.jacc.2019.09.056.

[88] Bryan Abadie et al. "Prevalence of ATTR-CA and high-risk features to guide testing in patients referred for TAVR." In: *European Journal of Nuclear Medicine and Molecular Imaging* 50.13 (Nov. 2023), pp. 3910–3916. DOI: 10.1007/s00259-023-06374-2.

[89] Pablo Garcia-Pavia et al. "Diagnosis and treatment of cardiac amyloidosis: a position statement of the ESC Working Group on Myocardial and Pericardial Diseases." en. In: *European Heart Journal* 42.16 (Apr. 2021), pp. 1554–1568. DOI: 10.1093/eurheartj/ehab072.

[90] Michelle M. Kittleson, Mathew S. Maurer, Amrut V. Ambardekar, Renee P. Bullock-Palmer, Patricia P. Chang, Howard J. Eisen, Ajith P. Nair, Jose Nativi-Nicolau, Frederick L. Ruberg, and On behalf of the American Heart Association Heart Failure and Transplantation Committee of the Council on Clinical Cardiology. "Cardiac Amyloidosis: Evolving Diagnosis and Management: A Scientific Statement From the American Heart Association." In: *Circulation* 142.1 (July 2020). DOI: 10.1161/CIR.0000000000000792.

[91] Karen Rausch, Gregory M. Scalia, Kei Sato, Natalie Edwards, Alfred King-yin Lam, David G. Platts, and Jonathan Chan. "Left atrial strain imaging differentiates cardiac amyloidosis and hypertensive heart disease." In: *The International Journal of Cardiovascular Imaging* 37.1 (Jan. 2021), pp. 81–90. DOI: 10.1007/s10554-020-01948-9.

[92] Jeremy A. Slivnick, Alexander L. Wallner, Ajay Vallakati, Vien T. Truong, Wojciech Mazur, Mohamed B. Elamin, Matthew S. Tong, Subha V. Raman, and Karolina M. Zareba. "Indexed left ventricular mass to QRS voltage ratio is associated with heart failure hospitalizations in patients with cardiac amyloidosis." In: *The International Journal of Cardiovascular Imaging* 37.3 (Mar. 2021), pp. 1043–1051. DOI: 10.1007/s10554-020-02059-1.

[93] María Del Carmen Mallón Araujo, Estephany Abou Jokh Casas, Charigan Abou Jokh Casas, María Amparo Martinez Monzonis, Álvaro Ruibal Morell, and Virginia Pubul Núñez. "Cardiac scintigraphy and echocardiography assessment in the diagnosis of transthyretin cardiac amyloidosis." In: *The Interna-*

tional Journal of Cardiovascular Imaging 40.2 (Nov. 2023), pp. 415–424. DOI: 10.1007/s10554-023-02987-8.

[94] H. S. A. Tingen, A. Tubben, J. H. Van 'T Oever, E. M. Pastoor, P. P. A. Van Zon, H. L. A. Nienhuis, P. Van Der Meer, and R. H J. A. Slart. "Positron emission tomography in the diagnosis and follow-up of transthyretin amyloid cardiomyopathy patients: A systematic review." In: *European Journal of Nuclear Medicine and Molecular Imaging* 51.1 (Dec. 2023), pp. 93–109. DOI: 10.1007/s00259-023-06381-3.

[95] María Del Carmen Navarro-Saez, Carlos Feijoo-Massó, Zully Del Carmen Bravo Ferrer, Joan Carles Oliva Morera, Andrea María Balado González, Alba Palau-Domínguez, Laura Guillamon Toran, Ricard Comet Monte, and Andreu Fernández-Codina. "Trends in diagnosis of cardiac transthyretin amyloidosis: 3-year analysis of scintigraphic studies: Prevalence of myocardial uptake and its predictor factors." In: *The International Journal of Cardiovascular Imaging* 39.7 (Apr. 2023), pp. 1397–1404. DOI: 10.1007/s10554-023-02840-y.

[96] Jolien Geers et al. "Prognostic value of left ventricular global constructive work in patients with cardiac amyloidosis." In: *The International Journal of Cardiovascular Imaging* 39.3 (Dec. 2022), pp. 585–593. DOI: 10.1007/s10554-022-02762-1.

[97] Christine P. Shen, Christopher T. Vanichsarn, Amitabh C. Pandey, Kristen Billick, David S. Rubenson, Rajeev C. Mohan, James Thomas Heywood, and Ajay V. Srivastava. "Wild type cardiac amyloidosis: is it time to order a nuclear technetium pyrophosphate SPECT imaging study?" In: *The International Journal of Cardiovascular Imaging* 39.1 (July 2022), pp. 201–208. DOI: 10.1007/s10554-022-02692-y.

[98] Nazim Coskun, M. Oguz Kartal, A. Sinem Erdogan, Omac Tufekcioglu, and Elif Ozdemir. "Tc-99m pyrophosphate scintigraphy for cardiac amyloidosis: concordance between planar and SPECT/CT imaging." In: *The International Journal of Cardiovascular Imaging* 38.9 (July 2022), pp. 2081–2088. DOI: 10.1007/s10554-022-02676-y.

[99] Selcan Kaplan Berkaya, Alper Kursat Uysal, Efnan Sora Gunal, Semih Ergin, Serkan Gunal, and M. Bilginer Gulmezoglu. "A survey on ECG analysis." In: *Biomedical Signal Processing and Control* 43 (May 2018), pp. 216–235. ISSN: 1746-8094. DOI: 10.1016/j.bspc.2018.03.003.

[100] Antoni Bayés De Luna, Velislav Batchvarov, and Marek Malik. "The morphology of the electrocardiogram." In: *The ESC Textbook of Cardiovascular Medicine*. Malden, Mass: Blackwell Pub, 2006. ISBN: 978-1-4051-2695-3.

[101] H. Y. Lin, S. Y. Liang, Y. L. Ho, Y. H. Lin, and H. P. Ma. "Discrete-wavelet-transform-based noise removal and feature extraction for ECG signals." In: *IRBM*. Healthcom 2013 35.6 (Dec. 2014), pp. 351–361. ISSN: 1959-0318. DOI: 10.1016/j.irbm.2014.10.004.

[102] Wilhelm Einthoven. "The different forms of the human electrocardiogram and their signification." In: *The Lancet*. Originally published as Volume 1, Issue 4622 179.4622 (Mar. 1912), pp. 853–861. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(00)50560-1.

[103] Antonio Bayés de Luna. "Introduction." In: *The ESC Textbook of Cardiovascular Medicine*. Ed. by A. John Camm, Thomas F. Lüscher, Gerald Maurer, and Patrick W. Serruys. 3rd ed. Oxford University Press, Dec. 2018, p. 0. ISBN: 978-0-19-878490-6.

[104] Ana Mincholé, Julià Camps, Aurore Lyon, and Blanca Rodríguez. "Machine learning in the electrocardiogram." In: *Journal of Electrocardiology* 57 (Nov. 2019), S61–S64. ISSN: 0022-0736. DOI: 10.1016/j.jelectrocard.2019.08.008.

[105] Sulaiman Somani et al. "Deep learning and the electrocardiogram: review of the current state-of-the-art." In: *EP Europace* 23.8 (Aug. 2021), pp. 1179–1191. ISSN: 1099-5129. DOI: 10.1093/europace/euaa377.

[106] Xinwen Liu, Huan Wang, Zongjin Li, and Lang Qin. "Deep learning in ECG diagnosis: A review." In: *Knowledge-Based Systems* 227 (Sept. 2021), p. 107187. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2021.107187.

[107] Koen Nieman. "Coronary computed tomography angiography: detection of coronary artery stenosis." In: *The ESC Textbook of Cardiovascular Medicine*. Ed. by A. John Camm, Thomas F. Lüscher, Gerald Maurer, Patrick W. Serruys, and Stephan Achenbach. Oxford University Press, Dec. 2018, p. 0. ISBN: 978-0-19-878490-6.

[108] Riccardo Marano, Giuseppe Rovere, Giancarlo Savino, Francesco Ciriaco Flammia, Maria Rachele Pia Carafa, Lorenzo Steri, Biagio Merlino, and Luigi Natale. "CCTA in the diagnosis of coronary artery disease." In: *La radiologia medica* 125.11 (Nov. 2020), pp. 1102–1113. ISSN: 1826-6983. DOI: 10.1007/s11547-020-01283-y.

[109] Konstantinos C. Siontis, Peter A. Noseworthy, Zachi I. Attia, and Paul A. Friedman. "Artificial intelligence-enhanced electrocardiography in cardiovascular disease management." In: *Nature Reviews Cardiology* 18.7 (July 2021), pp. 465–478. ISSN: 1759-5010. DOI: 10.1038/s41569-020-00503-2.

[110] Alvaro Alonso et al. "Simple Risk Model Predicts Incidence of Atrial Fibrillation in a Racially and Geographically Diverse Population: the CHARGE-AF Consortium." In: *Journal of the American Heart Association* 2.2 (Mar. 2013), e000102. DOI: 10.1161/JAHA.112.000102.

[111] Fons J. Wesselius, Mathijs S. van Schie, Natasja M. S. De Groot, and Richard C. Hendriks. "Digital biomarkers and algorithms for detection of atrial fibrillation using surface electrocardiograms: A systematic review." In: *Computers in Biology and Medicine* 133 (June 2021), p. 104404. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2021.104404.

[112] Zeineb Bouzid, Ziad Faramand, Richard E. Gregg, Stephanie Helman, Christian Martin-Gill, Samir Saba, Clifton Callaway, Ervin Sejdić, and Salah Al-Zaiti. "Novel ECG features and machine learning to optimize culprit lesion detection in patients with suspected acute coronary syndrome." In: *Journal of Electrocardiology* 69 (Nov. 2021), pp. 31–37. ISSN: 0022-0736. DOI: 10.1016/j.jelectrocard.2021.07.012.

[113] Stergios Intzes et al. "P-wave duration and atrial fibrillation recurrence after catheter ablation: a systematic review and meta-analysis." In: *EP Europace* 25.2 (Feb. 2023), pp. 450–459. ISSN: 1099-5129. DOI: 10.1093/europace/euac210.

[114] Claudia Nagel, Giorgio Luongo, Luca Azzolin, Steffen Schuler, Olaf Dössel, and Axel Loewe. "Non-Invasive and Quantitative Estimation of Left Atrial Fibrosis Based on P Waves of the 12-Lead ECG—A Large-Scale Computational Study Covering Anatomical Variability." In: *Journal of Clinical Medicine* 10.8 (Jan. 2021), p. 1797. ISSN: 2077-0383. DOI: 10.3390/jcm10081797.

[115] Antonio Luiz P. Ribeiro et al. "Tele-electrocardiography and bigdata: The CODE (Clinical Outcomes in Digital Electrocardiography) study." In: *Journal of Electrocardiology* 57 (Nov. 2019), S75–S78. ISSN: 0022-0736. DOI: 10.1016/j.jelectrocard.2019.09.008.

[116] Ben Van Calster et al. "Calibration: the Achilles heel of predictive analytics." In: *BMC Medicine* 17.1 (Dec. 2019), p. 230. ISSN: 1741-7015. DOI: 10.1186/s12916-019-1466-7.

[117] Yingxiang Huang, Wentao Li, Fima Macheret, Rodney A Gabriel, and Lucila Ohno-Machado. "A tutorial on calibration measurements and calibration models for clinical prediction models." In: *Journal of the American Medical Informatics Association* 27.4 (Apr. 2020), pp. 621–633. ISSN: 1527-974X. DOI: 10.1093/jamia/ocz228.

[118] Ben Van Calster and Andrew J. Vickers. "Calibration of Risk Prediction Models: Impact on Decision-Analytic Performance." In: *Medical Decision Making* 35.2 (Feb. 2015), pp. 162–169. ISSN: 0272-989X. DOI: 10.1177/0272989X14547233.

[119] D. R. COX. "Two further applications of a model for binary regression." In: *Biometrika* 45.3-4 (Dec. 1958), pp. 562–565. ISSN: 0006-3444. DOI: 10.1093/biomet/45.3-4.562.

[120] Glenn W Brier. "Verification of forecasts expressed in terms of probability." In: *Monthly weather review* 78.1 (1950), pp. 1–3.

[121] David W. Hosmer and Stanley Lemesbow. "Goodness of fit tests for the multiple logistic regression model." In: *Communications in Statistics - Theory and Methods* 9.10 (Jan. 1980), pp. 1043–1069. ISSN: 0361-0926. DOI: 10.1080/03610928008827941.

[122] Peter C. Austin and Ewout W. Steyerberg. "The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models." In: *Statistics in Medicine* 38.21 (2019), pp. 4051–4065. ISSN: 1097-0258. DOI: 10.1002/sim.8281.

[123] Haibo He and Edwardo A. Garcia. "Learning from Imbalanced Data." In: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (Sept. 2009), pp. 1263–1284. ISSN: 1558-2191. DOI: 10.1109/TKDE.2008.239.

[124] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from Imbalanced Data Sets*. Cham: Springer International Publishing, 2018. ISBN: 978-3-319-98073-7. DOI: 10.1007/978-3-319-98074-4.

[125] Fadel M. Megahed, Ying-Ju Chen, Aly Megahed, Yuya Ong, Naomi Altman, and Martin Krzywinski. "The class imbalance problem." In: *Nature Methods* 18.11 (Nov. 2021), pp. 1270–1272. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01302-4.

[126] Ruben van den Goorbergh, Maarten van Smeden, Dirk Timmerman, and Ben Van Calster. "The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression." In: *Journal of the American Medical Informatics Association* 29.9 (Sept. 2022), pp. 1525–1534. ISSN: 1527-974X. DOI: 10.1093/jamia/ocac093.

[127] Annamaria Iorio, Gianfranco Sinagra, and Andrea Di Lenarda. "Administrative database, observational research and the Tower of Babel." In: *International Journal of Cardiology* 284 (June 2019), pp. 118–119. ISSN: 0167-5273, 1874-1754. DOI: 10.1016/j.ijcard.2018.12.009.

[128] Giulia Barbati, Francesca Ieva, Francesca Gasperoni, Annamaria Iorio, Gianfranco Sinagra, and Andrea Di Lenarda. "The Trieste Observatory of cardiovascular disease: an experience of administrative and clinical data integration at a regional level." In: *SIS 2017. Statistics and Data Science: new challenges, new generations*. Florence: Firenze University Press, 2017, pp. 115–122. DOI: 10.36253/978-88-6453-521-0.

[129] A. Scagnetto, G. Barbati, I. Gandin, C. Cappelletto, G. Baj, A. Cazzaniga, F. Cuturello, A. Ansuini, L. Bortolussi, and A. Di Lenarda. *Deep artificial neural network for prediction of atrial fibrillation through the analysis of 12-leads standard ECG*. arXiv:2202.05676 [cs, eess]. Jan. 2022. DOI: 10.48550/arXiv.2202.05676.

[130] Sebastian D. Goodfellow, Andrew Goodwin, Robert Greer, Peter C. Laussen, Mjaye Mazwi, and Danny Eytan. "Towards Understanding ECG Rhythm Classification Using Convolutional Neural Networks and Attention Mappings." In: *Proceedings of the 3rd Machine Learning for Healthcare Conference*. PMLR, Nov. 2018, pp. 83–101.

[131] Ilya Loshchilov and Frank Hutter. *Decoupled Weight Decay Regularization*. arXiv:1711.05101 [cs, math]. Jan. 2019. DOI: 10.48550/arXiv.1711.05101.

[132] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.

[133] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python." In: *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830. ISSN: 1533-7928.

[134] Laila Staerk, Jason A. Sherer, Darae Ko, Emelia J. Benjamin, and Robert H. Helm. "Atrial Fibrillation." In: *Circulation Research* 120.9 (Apr. 2017), pp. 1501–1517. DOI: 10.1161/CIRCRESAHA.117.309732.

[135] Antonio Di Carlo et al. "Prevalence of atrial fibrillation in the Italian elderly population and projections from 2020 to 2060 for Italy and the European Union: the FAI Project." In: *EP Europace* 21.10 (Oct. 2019), pp. 1468–1475. DOI: 10.1093/europace/euz141.

[136] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. "On calibration of modern neural networks." In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, Aug. 2017, pp. 1321–1330.

[137] Sherri Rose. "Machine Learning for Prediction in Electronic Health Data." In: *JAMA Network Open* 1.4 (Aug. 2018), e181404. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2018.1404.

[138] Shaan Khurshid et al. "ECG-Based Deep Learning and Clinical Risk Factors to Predict Atrial Fibrillation." In: *Circulation* 145.2 (Jan. 2022), pp. 122–133. DOI: 10.1161/CIRCULATIONAHA.121.057480.

[139] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." In: *J. Artif. Int. Res.* 16.1 (June 2002), pp. 321–357. ISSN: 1076-9757.

[140] Andres Hernandez-Matamoros, Hamido Fujita, and Hector Perez-Meana. "A novel approach to create synthetic biomedical signals using BiRNN." In: *Information Sciences* 541 (Dec. 2020), pp. 218–241. ISSN: 0020-0255. DOI: 10.1016/j.ins.2020.06.019.

[141] Fei Zhu, Fei Ye, Yuchen Fu, Quan Liu, and Bairong Shen. "Electrocardiogram generation with a bidirectional LSTM-CNN generative adversarial network." In: *Scientific Reports* 9.1 (May 2019), p. 6734. ISSN: 2045-2322. DOI: 10.1038/s41598-019-42516-z.

[142] Edmond Adib, Fatemeh Afghah, and John J. Prevost. *Synthetic ECG Signal Generation Using Generative Neural Networks*. Aug. 2022. DOI: 10.48550/arXiv.2112.03268.

[143] Karli Gillette et al. "MedalCare-XL: 16,900 healthy and pathological synthetic 12 lead ECGs from electrophysiological simulations." In: *Scientific Data* 10.1 (Aug. 2023), p. 531. ISSN: 2052-4463. DOI: 10.1038/s41597-023-02416-4.

[144] Shany Biton, Sheina Gendelman, Antônio H Ribeiro, Gabriela Miana, Carla Moreira, Antonio Luiz P Ribeiro, and Joachim A Behar. "Atrial fibrillation risk prediction from the 12-lead electrocardiogram using digital biomarkers and deep representation learning." In: *European Heart Journal - Digital Health* 2.4 (Dec. 2021), pp. 576–585. ISSN: 2634-3916. DOI: 10.1093/ehjdh/ztab071.

[145] Björn Müller-Edenborn, Juan Chen, Jürgen Allgeier, Maxim Didenko, Zoraida Moreno-Weidmann, Franz-Josef Neumann, Heiko Lehrmann, Reinhold Weber, Thomas Arentz, and Amir Jadidi. "Amplified sinus-P-wave reveals localization and extent of left atrial low-voltage substrate: implications for arrhythmia freedom following pulmonary vein isolation." In: *EP Europace* 22.2 (Feb. 2020), pp. 240–249. ISSN: 1099-5129. DOI: 10.1093/europace/euz297.

[146] Giulia Barbati et al. "Study Design and Research Protocol for diagnostic or prognostic studies in the Age of Artificial Intelligence: A Biostatistician's Perspective." In: *Epidemiology, Biostatistics, and Public Health* 18.2 (2023). DOI: 10.54103/2282-

0930/22227. URL: https://riviste.unimi.it/index.php/ebph/article/view/22227.

[147] Samantha Cruz Rivera, Xiaoxuan Liu, An-Wen Chan, Alastair K. Denniston, and Melanie J. Calvert. "Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension." In: *Nature Medicine* 26.9 (Sept. 2020), pp. 1351–1363. DOI: 10.1038/s41591-020-1037-7.

[148] Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J. Calvert, and Alastair K. Denniston. "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension." In: *Nature Medicine* 26.9 (Sept. 2020), pp. 1364–1374. DOI: 10.1038/s41591-020-1034-x.

[149] B. Vasey, A. Novak, S. Ather, M. Ibrahim, and P. McCulloch. "DECIDE-AI: a new reporting guideline and its relevance to artificial intelligence studies in radiology." In: *Clinical Radiology*. Special Issue Section: Artificial Intelligence and Machine Learning 78.2 (Feb. 2023), pp. 130–136. DOI: 10.1016/j.crad.2022.09.131.

[150] Gary S Collins et al. "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods." In: *BMJ* (Apr. 2024), e078378. DOI: 10.1136/bmj-2023-078378.

[151] Tjeerd Van Der Ploeg, Peter C Austin, and Ewout W Steyerberg. "Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints." In: *BMC Medical Research Methodology* 14.1 (Dec. 2014), p. 137. DOI: 10.1186/1471-2288-14-137.

[152] Ahmad Alwosheel, Sander Van Cranenburgh, and Caspar G. Chorus. "Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis." In: *Journal of Choice Modelling* 28 (Sept. 2018), pp. 167–182. DOI: 10.1016/j.jocm.2018.07.002.

[153] Gabriele Infante, Rosalba Miceli, and Federico Ambrogi. "Sample size and predictive performance of machine learning methods with survival data: A simulation study." In: *Statistics in Medicine* 42.30 (Dec. 2023), pp. 5657–5675. DOI: 10.1002/sim.9931.

[154] Sayan Mukherjee, Pablo Tamayo, Simon Rogers, Ryan Rifkin, Anna Engle, Colin Campbell, Todd R. Golub, and Jill P. Mesirov. "Estimating Dataset Size Requirements for Classifying DNA Microarray Data." In: *Journal of Computational Biology* 10.2 (Apr. 2003), pp. 119–142. DOI: 10.1089/106652703321825928.

[155] Rosa L. Figueroa, Qing Zeng-Treitler, Sasikiran Kandula, and Long H. Ngo. "Predicting sample size required for classification performance." In: *BMC Medical Informatics and Decision Making* 12.1 (Feb. 2012), p. 8. ISSN: 1472-6947. DOI: 10.1186/1472-6947-12-8.

[156] Alimu Dayimu, Nikola Simidjievski, Nikolaos Demiris, and Jean Abraham. "Sample size determination for prediction models via learning-type curves." In: *Statistics in Medicine* 43.16 (2024), pp. 3062–3072. ISSN: 1097-0258. DOI: 10.1002/sim.10121.

[157] David Jean Biau, Solen Kernéis, and Raphaël Porcher. "Statistics in Brief: The Importance of Sample Size in the Planning and Interpretation of Medical Research." In: *Clinical Orthopaedics and Related Research* 466.9 (Sept. 2008), pp. 2282–2288. DOI: 10.1007/s11999-008-0346-9.

[158] Indranil Balki et al. "Sample-Size Determination Methodologies for Machine Learning in Medical Imaging Research: A Systematic Review." In: *Canadian Association of Radiologists Journal* 70.4 (Nov. 2019), pp. 344–353. DOI: 10.1016/j.carj.2019.06.002.

[159] Tom Viering and Marco Loog. "The Shape of Learning Curves: A Review." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.6 (June 2023), pp. 7799–7819. DOI: 10.1109/TPAMI.2022.3220744.

[160] Erick A. Perez Alday et al. "Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020." In: *Physiological Measurement* 41.12 (Dec. 2020), p. 124003. DOI: 10.1088/1361-6579/abc960.

[161] Georgios Christopoulos et al. "Artificial Intelligence–Electrocardiography to Predict Incident Atrial Fibrillation." In: *Circulation: Arrhythmia and Electrophysiology* 13.12 (Dec. 2020), e009355. DOI: 10.1161/CIRCEP.120.009355.

[162] Peter C. Austin, Douglas S. Lee, and Jason P. Fine. "Introduction to the Analysis of Survival Data in the Presence of Competing Risks." In: *Circulation* 133.6 (Feb. 2016), pp. 601–609. DOI: 10.1161/CIRCULATIONAHA.115.017719.

[163] Laura Lee Johnson. "Chapter 26 - An Introduction to Survival Analysis." In: *Principles and Practice of Clinical Research*. Ed. by John I. Gallin, Frederick P. Ognibene, and Laura Lee Johnson. 4th ed. Boston: Academic Press, Jan. 2018, pp. 373–381. ISBN: 978-0-12-849905-4.

[164] David G. Kleinbaum and Mitchel Klein. *Survival Analysis: A Self-Learning Text*. Statistics for Biology and Health. New York, NY: Springer, 2012. ISBN: 978-1-4419-6645-2 978-1-4419-6646-9. DOI: 10.1007/978-1-4419-6646-9.

[165] E. L. Kaplan and Paul Meier. "Nonparametric Estimation from Incomplete Observations." In: *Journal of the American Statistical Association* 53.282 (June 1958), pp. 457–481. ISSN: 0162-1459. DOI: 10.1080/01621459.1958.10501452.

[166] Wayne Nelson. "Hazard Plotting for Incomplete Failure Data." In: *Journal of Quality Technology* 1.1 (Jan. 1969), pp. 27–52. DOI: 10.1080/00224065.1969.11980344.

[167] Wayne Nelson. "Theory and Applications of Hazard Plotting for Censored Failure Data." In: *Technometrics* 14.4 (Nov. 1972), pp. 945–966. ISSN: 0040-1706. DOI: 10.1080/00401706.1972.10488991.

[168] Odd Aalen. "Nonparametric Inference for a Family of Counting Processes." In: *The Annals of Statistics* 6.4 (July 1978), pp. 701–726. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1176344247.

[169] D. R. Cox. "Regression Models and Life-Tables." In: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2 (1972), pp. 187–220. ISSN: 0035-9246. URL: https://www.jstor.org/stable/2985181 (visited on 08/23/2024).

[170] Husam Abdel-Qadir, Jiming Fang, Douglas S. Lee, Jack V. Tu, Eitan Amir, Peter C. Austin, and Geoffrey M. Anderson. "Importance of Considering Competing Risks in Time-to-Event Analyses." In: *Circulation: Cardiovascular Quality and Outcomes* 11.7 (July 2018), e004580. DOI: 10.1161/CIRCOUTCOMES.118.004580.

[171] H. Putter, M. Fiocco, and R. B. Geskus. "Tutorial in biostatistics: competing risks and multi-state models." In: *Statistics in Medicine* 26.11 (2007), pp. 2389–2430. ISSN: 1097-0258. DOI: 10.1002/sim.2712. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2712 (visited on 08/23/2024).

[172] Jason P. Fine and Robert J. Gray. "A Proportional Hazards Model for the Subdistribution of a Competing Risk." In: *Journal of the American Statistical Association* 94.446 (June 1999), pp. 496–509. DOI: 10.1080/01621459.1999.10474144.

[173] Ping Wang, Yan Li, and Chandan K. Reddy. "Machine Learning for Survival Analysis: A Survey." In: *ACM Comput. Surv.* 51.6 (Feb. 2019), 110:1–110:36. ISSN: 0360-0300. DOI: 10.1145/3214306.

[174] Imad Bou-Hamad, Denis Larocque, and Hatem Ben-Ameur. "A review of survival trees." In: *Statistics Surveys* 5.none (Jan. 2011), pp. 44–71. ISSN: 1935-7516. DOI: 10.1214/09-SS047.

[175] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. "Random survival forests." In: *The Annals of Applied Statistics* 2.3 (Sept. 2008), pp. 841–860. ISSN: 1932-6157, 1941-7330. DOI: 10.1214/08-AOAS169.

[176] Harald Binder and Martin Schumacher. "Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models." In: *BMC Bioinformatics* 9.1 (Jan. 2008), p. 14. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-14.

[177] Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. "Deep Survival Analysis." In: *Proceedings of the 1st Machine Learning for Healthcare Conference*. PMLR, Dec. 2016, pp. 101–114. URL: https://proceedings.mlr.press/v56/Ranganath16.html.

[178] Sebastian Pölsterl, Nassir Navab, and Amin Katouzian. "Fast Training of Support Vector Machines for Survival Analysis." In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Annalisa Appice, Pedro Pereira Rodrigues, Vítor Santos Costa, João Gama, Alípio Jorge, and Carlos Soares. Cham: Springer International Publishing, 2015, pp. 243–259. ISBN: 978-3-319-23525-7. DOI: 10.1007/978-3-319-23525-7_15.

[179] Tamara Fernandez, Nicolas Rivera, and Yee Whye Teh. "Gaussian Processes for Survival Analysis." In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc., 2016. URL: https://papers.nips.cc/paper_files/paper/2016/hash/ef1e491a766ce3127556063d49bc2f98-Abstract.html.

[180] David Faraggi and Richard Simon. "A neural network model for survival data." In: *Statistics in Medicine* 14.1 (1995), pp. 73–82. ISSN: 1097-0258. DOI: 10.1002/sim.4780140108.

[181] Simon Wiegrebe, Philipp Kopper, Raphael Sonabend, Bernd Bischl, and Andreas Bender. "Deep learning for survival analysis: a review." In: *Artificial Intelligence Review* 57.3 (Feb. 2024), p. 65. ISSN: 1573-7462. DOI: 10.1007/s10462-023-10681-3.

[182] Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network." In: *BMC Medical Research Methodology* 18.1 (Feb. 2018), p. 24. ISSN: 1471-2288. DOI: 10.1186/s12874-018-0482-1.

[183] Travers Ching, Xun Zhu, and Lana X. Garmire. "Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data." In: *PLOS Computational Biology* 14.4 (Apr. 2018), e1006076. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006076.

[184] Xinliang Zhu, Jiawen Yao, and Junzhou Huang. "Deep convolutional neural network for survival analysis with pathological images." In: *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Dec. 2016, pp. 544–547. DOI: 10.1109/BIBM.2016.7822579.

[185] Håvard Kvamme and Ørnulf Borgan. "Continuous and discrete-time survival prediction with neural networks." In: *Lifetime Data Analysis* 27.4 (Oct. 2021), pp. 710–736. ISSN: 1572-9249. DOI: 10.1007/s10985-021-09532-6.

[186] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. "DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (Apr. 2018). ISSN: 2374-3468. DOI: 10.1609/aaai.v32i1.11842.

[187] Michael F. Gensheimer and Balasubramanian Narasimhan. "A scalable discrete-time survival model for neural networks." In: *PeerJ* 7 (Jan. 2019), e6257. ISSN: 2167-8359. DOI: 10.7717/peerj.6257.

[188] Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P. Lungren. "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines." In: *npj Digital Medicine* 3.1 (Oct. 2020), pp. 1–9. ISSN: 2398-6352. DOI: 10.1038/s41746-020-00341-z.

[189] Paul Kligfield et al. "Recommendations for the Standardization and Interpretation of the Electrocardiogram." In: *Circulation* 115.10 (Mar. 2007), pp. 1306–1324. DOI: 10.1161/CIRCULATIONAHA.106.180200.

[190] Adina Najwa Kamarudin, Trevor Cox, and Ruwanthi Kolamunnage-Dona. "Time-dependent ROC curve analysis in medical research: current methods and applications." In: *BMC Medical Research Methodology* 17.1 (Apr. 2017), p. 53. ISSN: 1471-2288. DOI: 10.1186/s12874-017-0332-6.

[191] Peter C. Austin, Hein Putter, Daniele Giardiello, and David van Klaveren. "Graphical calibration curves and the integrated calibration index (ICI) for competing risk models." In: *Diagnostic and Prognostic Research* 6.1 (Jan. 2022), p. 2. ISSN: 2397-7523. DOI: 10.1186/s41512-021-00114-6.

[192] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M. Hospedales. "Self-Supervised Representation Learning: Introduction, advances, and challenges." In: *IEEE Signal Processing Magazine* 39.3 (May 2022), pp. 42–62. ISSN: 1558-0792. DOI: 10.1109/MSP.2021.3134634.

[193] Temesgen Mehari and Nils Strodthoff. "Self-supervised representation learning from 12-lead ECG data." In: *Computers in Biology and Medicine* 141 (Feb. 2022), p. 105114. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2021.105114.

[194] Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela van der Schaar. "VIME: Extending the Success of Self- and Semi-supervised Learning to Tabular Domain." In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 11033–11043.

[195] Yongshuo Zong, Oisin Mac Aodha, and Timothy Hospedales. *Self-Supervised Multimodal Learning: A Survey*. Aug. 2024. DOI: 10.48550/arXiv.2304.01008.

[196] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan E. Taylor. *An introduction to statistical learning: with applications in Python*. Springer texts in statistics. Cham, Switzerland: Springer, 2023. ISBN: 978-3-031-38746-3 978-3-031-39189-7.

[197] Abiodun M. Ikotun, Absalom E. Ezugwu, Laith Abualigah, Belal Abuhaija, and Jia Heming. "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data." In: *Information Sciences* 622 (Apr. 2023), pp. 178–210. DOI: 10.1016/j.ins.2022.11.139.

[198] Xingcheng Ran, Yue Xi, Yonggang Lu, Xiangwen Wang, and Zhenyu Lu. "Comprehensive survey on hierarchical clustering algorithms and the recent developments." In: *Artificial Intelligence Review* 56.8 (Aug. 2023), pp. 8219–8264. ISSN: 1573-7462. DOI: 10.1007/s10462-022-10366-3.

[199] Thaddeus Tarpey and Kimberly K. J. Kinateder. "Clustering Functional Data." In: *Journal of Classification* 20.1 (May 2003), pp. 093–114. ISSN: 1432-1343. DOI: 10.1007/s00357-003-0007-3.

[200] Mimi Zhang and Andrew Parnell. "Review of Clustering Methods for Functional Data." In: *ACM Trans. Knowl. Discov. Data* 17.7 (Apr. 2023), 91:1–91:34. ISSN: 1556-4681. DOI: 10.1145/3581789.

[201] Alessia Paldino et al. "Prognostic Prediction of Genotype vs Phenotype in Genetic Cardiomyopathies." In: *Journal of the American College of Cardiology* 80.21 (Nov. 2022), pp. 1981–1994. DOI: 10.1016/j.jacc.2022.08.804.

[202] Jesús G. Mirelis et al. "Combination of late gadolinium enhancement and genotype improves prediction of prognosis in non-ischaemic dilated cardiomyopathy." In: *European Journal of Heart Failure* 24.7 (2022), pp. 1183–1196. DOI: 10.1002/ejhf.2514.

[203] Marta Gigli et al. "Genetic Risk of Arrhythmic Phenotypes in Patients With Dilated Cardiomyopathy." In: *Journal of the American College of Cardiology* 74.11 (Sept. 2019), pp. 1480–1490. DOI: 10.1016/j.jacc.2019.06.072.

[204] Ewa Dziewięcka, Matylda Gliniak, Mateusz Winiarczyk, Arman Karapetyan, Sylwia Wiśniowska-Śmiałek, Aleksandra Karabinowska, Marcin Dziewięcki, Piotr Podolec, and Paweł Rubiś. "Mortality risk in dilated cardiomyopathy: the accuracy of heart failure prognostic models and dilated cardiomyopathy-tailored prognostic model." In: *ESC Heart Failure* 7.5 (2020), pp. 2455–2467. DOI: 10.1002/ehf2.12809.

[205] Upasana Tayal et al. "Precision Phenotyping of Dilated Cardiomyopathy Using Multidimensional Data." In: *Journal of the American College of Cardiology* 79.22 (June 2022), pp. 2219–2232. DOI: 10.1016/j.jacc.2022.03.375.

[206] Job A. J. Verdonschot et al. "Clustering of Cardiac Transcriptome Profiles Reveals Unique: Subgroups of Dilated Cardiomyopathy Patients." In: *JACC: Basic to Translational Science* 8.4 (Apr. 2023), pp. 406–418. DOI: 10.1016/j.jacbts.2022.10.010.

[207] Lopez Luis Escobar et al. "Clinical Risk Score to Predict Pathogenic Genotypes in Patients With Dilated Cardiomyopathy." In: *Journal of the American College of Cardiology* 80.12 (Sept. 2022), pp. 1115–1126. DOI: 10.1016/j.jacc.2022.06.040.

[208] Marta Gigli et al. "Phenotypic Expression, Natural History, and Risk Stratification of Cardiomyopathy Caused by Filamin C Truncating Variants." In: *Circulation* 144.20 (Nov. 2021), pp. 1600–1611. DOI: 10.1161/CIRCULATIONAHA.121.053521.

[209] Elena Arbelo et al. "2023 ESC Guidelines for the management of cardiomyopathies: Developed by the task force on the management of cardiomyopathies of the European Society of Cardiology (ESC)." In: *European Heart Journal* 44.37 (Oct. 2023), pp. 3503–3626. DOI: 10.1093/eurheartj/ehad194.

[210] Roberto M. Lang et al. "Recommendations for Cardiac Chamber Quantification by Echocardiography in Adults: An Update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging." In: *European Heart Journal - Cardiovascular Imaging* 16.3 (Mar. 2015), pp. 233–271. DOI: 10.1093/ehjci/jev014.

[211] Matteo Dal Ferro et al. "Association between mutation status and left ventricular reverse remodelling in dilated cardiomyopathy." In: *Heart* 103.21 (Nov. 2017), pp. 1704–1710. DOI: 10.1136/heartjnl-2016-311017.

[212] Sue Richards et al. "Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology." In: *Genetics in Medicine* 17.5 (May 2015), pp. 405–423.

[213] Peter J. Rousseeuw. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis." In: *Journal of Computational and Applied Mathematics* 20 (Nov. 1987), pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

[214] Robert Tibshirani. "Regression Shrinkage and Selection Via the Lasso." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (Jan. 1996), pp. 267–288. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

[215] Marco Merlo, Denise Zaffalon, Davide Stolfo, Alessandro Altinier, Giulia Barbati, Massimo Zecchin, Stefano Bardari, and Gianfranco Sinagra. "ECG in dilated cardiomyopathy: specific findings and long-term prognostic significance." In: *Journal of Cardiovascular Medicine* 20.7 (July 2019), p. 450. DOI: 10.2459/JCM.0000000000000804.

[216] Patrick Royston and Mahesh K. B. Parmar. "Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects." In: *Statistics in Medicine* 21.15 (2002), pp. 2175–2197. DOI: 10.1002/sim.1203.

[217] Micha Mandel. "Simulation-Based Confidence Intervals for Functions With Complicated Derivatives." In: *The American Statistician* 67.2 (May 2013), pp. 76–81. DOI: 10.1080/00031305.2013.783880.

[218] Aymeric Stamm, Laura Sangalli, Piercesare Secchi, Simone Vantini, Valeria Vitelli, and Alessandro Zito. *fdacluster: Joint Clustering and Alignment of Functional Data*. July 2023. URL: https://cran.r-project.org/web/packages/fdacluster/index.html.

[219] Alexander Strehl and Joydeep Ghosh. "Relationship-Based Clustering and Visualization for High-Dimensional Data Mining." In: *INFORMS J. on Computing* 15.2 (Apr. 2003), pp. 208–230. ISSN: 1526-5528. DOI: 10.1287/ijoc.15.2.208.14448.

[220] A. D. Gordon and M. Vichi. "Fuzzy partition models for fitting a set of partitions." In: *Psychometrika* 66.2 (June 2001), pp. 229–247. ISSN: 1860-0980. DOI: 10.1007/BF02294837.

[221] Sara López-Pintado and Juan Romo. "On the Concept of Depth for Functional Data." In: *Journal of the American Statistical Association* 104.486 (June 2009), pp. 718–734. ISSN: 0162-1459. DOI: 10.1198/jasa.2009.0108.

[222] Francesca Ieva, Anna Maria Paganoni, Juan Romo, and Nicholas Tarabelloni. "roahd Package: Robust Analysis of High Dimensional Data." In: *The R Journal* 11.2 (2019), pp. 291–307. ISSN: 2073-4859. URL: https://journal.r-project.org/archive/2019/RJ-2019-032/index.html.

[223] Christoph Gräni, Marius R. Bigler, and Raymond Y. Kwong. "Noninvasive Multimodality Imaging for the Assessment of Anomalous Coronary Artery." In: *Current Cardiology Reports* 25.10 (Oct. 2023), pp. 1233–1246. ISSN: 1534-3170. DOI: 10.1007/s11886-023-01948-w.

[224] Marius Reto Bigler, Alexander Kadner, Lorenz Räber, Afreed Ashraf, Stephan Windecker, Matthias Siepe, Massimo Antonio Padalino, and Christoph Gräni. "Therapeutic Management of Anomalous Coronary Arteries Originating From the Opposite Sinus of Valsalva: Current Evidence, Proposed Approach, and the Unknowing." In: *Journal of the American Heart Association* 11.20 (Oct. 2022), e027098. DOI: 10.1161/JAHA.122.027098.

[225] Christoph Gräni, Ronny R. Buechel, Philipp A. Kaufmann, and Raymond Y. Kwong. "Multimodality Imaging in Individuals With Anomalous Coronary Arteries." In: *JACC: Cardiovascular Imaging* 10.4 (Apr. 2017), pp. 471–481. ISSN: 1936-878X. DOI: 10.1016/j.jcmg.2017.02.004.

[226] Juhani Knuuti et al. "2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromes: The Task Force for the diagnosis and management of chronic coronary syndromes of the European Society of Cardiology (ESC)." In: *European Heart Journal* 41.3 (Jan. 2020), pp. 407–477. DOI: 10.1093/eurheartj/ehz425.

[227] Karen K. Stout et al. "2018 AHA/ACC Guideline for the Management of Adults With Congenital Heart Disease: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines." In: *Journal of the American College of Cardiology* 73.12 (Apr. 2019), pp. 1494–1563. DOI: 10.1016/j.jacc.2018.08.1028.

[228] Paolo Angelini. "Coronary Artery Anomalies." In: *Circulation* 115.10 (Mar. 2007), pp. 1296–1305. DOI: 10.1161/CIRCULATIONAHA.106.618082.

[229] Geert Litjens, Francesco Ciompi, Jelmer M. Wolterink, Bob D. de Vos, Tim Leiner, Jonas Teuwen, and Ivana Išgum. "State-of-the-Art Deep Learning in Cardiovascular Image Analysis." In: *JACC: Cardiovascular Imaging* 12.8, Part 1 (Aug. 2019), pp. 1549–1565. DOI: 10.1016/j.jcmg.2019.06.009.

[230] Damini Dey et al. "Proceedings of the NHLBI Workshop on Artificial Intelligence in Cardiovascular Imaging: Translation to Patient Care." In: *JACC: Cardiovascular Imaging* 16.9 (Sept. 2023), pp. 1209–1223. ISSN: 1936-878X. DOI: 10.1016/j.jcmg.2023.05.012.

[231] Ali S. Tejani et al. "Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update." In: *Radiology: Artificial Intelligence* 6.4 (July 2024), e240300. DOI: 10.1148/ryai.240300.

[232] Viknesh Sounderajah et al. "Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol." In: *BMJ Open* 11.6 (June 2021), e047709. DOI: 10.1136/bmjopen-2020-047709.

[233] Tina Hernandez-Boussard, Selen Bozkurt, John P A Ioannidis, and Nigam H Shah. "MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care." In: *Journal of the American Medical Informatics Association* 27.12 (Dec. 2020), pp. 2011–2015. DOI: 10.1093/jamia/ocaa088.

[234] Jakob Wasserthal et al. "TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images." In: *Radiology: Artificial Intelligence* 5.5 (Sept. 2023), e230024. DOI: 10.1148/ryai.230024.

[235] An Zeng et al. "ImageCAS: A large-scale dataset and benchmark for coronary artery segmentation based on computed tomography angiography images." In: *Computerized Medical Imaging and Graphics* 109 (Oct. 2023), p. 102287. DOI: 10.1016/j.compmedimag.2023.102287.

[236] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation." In: *Nature Methods* 18.2 (Feb. 2021), pp. 203–211. DOI: 10.1038/s41592-020-01008-z.

[237] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-Excitation Networks." In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. June 2018, pp. 7132–7141. DOI: 10.1109/CVPR.2018.00745.

[238] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks." In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Mar. 2018, pp. 839–847. DOI: 10.1109/WACV.2018.00097.

[239] Barry J. Maron, Tammy S. Haas, Aneesha Ahluwalia, Caleb J. Murphy, and Ross F. Garberich. "Demographics and Epidemiology of Sudden Deaths in Young Competitive Athletes: From the United States National Registry." In: *The American Journal of Medicine* 129.11 (Nov. 2016), pp. 1170–1177. DOI: 10.1016/j.amjmed.2016.02.031.

[240] Paolo Angelini, Benjamin Y. Cheong, Veronica V. Lenge De Rosen, Alberto Lopez, Carlo Uribe, Anthony H. Masso, Syed W. Ali, Barry R. Davis, Raja Muthupillai, and James T. Willerson. "High-Risk Cardiovascular Conditions in Sports-Related Sudden Death: Prevalence in 5,169 Schoolchildren Screened via Cardiac Magnetic Resonance." In: *Texas Heart Institute Journal* 45.4 (Aug. 2018), pp. 205–213. DOI: 10.14503/THIJ-18-6645.

[241] Christoph Gräni et al. "Prevalence and characteristics of coronary artery anomalies detected by coronary computed tomography angiography in 5 634 consecutive patients in a single centre in Switzerland." In: *Swiss Medical Weekly* 146.1718 (Apr. 2016), w14294–w14294. ISSN: 1424-3997. DOI: 10.4414/smw.2016.14294.

[242] Jelena R. Ghadri et al. "Congenital coronary anomalies detected by coronary computed tomography compared to invasive coronary angiography." In: *BMC Cardiovascular Disorders* 14.1 (July 2014), p. 81. DOI: 10.1186/1471-2261-14-81.

[243] Francesco Gentile, Vincenzo Castiglione, and Raffaele De Caterina. "Coronary Artery Anomalies." In: *Circulation* 144.12 (Sept. 2021), pp. 983–996. DOI: 10.1161/CIRCULATIONAHA.121.055347.

[244] Christoph Gräni et al. "Outcome in middle-aged individuals with anomalous origin of the coronary artery from the opposite sinus: a matched cohort study." In: *European Heart Journal* 38.25 (July 2017), pp. 2009–2016. DOI: 10.1093/eurheartj/ehx046.

[245] Alexander R. van Rosendael et al. "Rationale and design of the CONFIRM2 (Quantitative COroNary CT Angiography Evaluation For Evaluation of Clinical Outcomes: An InteRnational, Multicenter Registry) study." In: *Journal of Cardiovascular Computed Tomography* 18.1 (Jan. 2024), pp. 11–17. DOI: 10.1016/j.jcct.2023.10.004.

[246] Ariel Fernando Pascaner, Antonio Rosato, Alice Fantazzini, Elena Vincenzi, Curzio Basso, Francesco Secchi, Mauro Lo Rito, and Michele Conti. "Automatic 3D Segmentation and Identification of Anomalous Aortic Origin of the Coronary Arteries Combining Multi-view 2D Convolutional Neural Networks." In: *Journal of Imaging Informatics in Medicine* 37.2 (Apr. 2024), pp. 884–891. DOI: 10.1007/s10278-023-00950-6.

[247] George C. M. Siontis, Ioanna Tzoulaki, Peter J. Castaldi, and John P. A. Ioannidis. "External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination." In: *Journal of Clinical Epidemiology* 68.1 (Jan. 2015), pp. 25–34. DOI: 10.1016/j.jclinepi.2014.09.007.

[248] George C. M. Siontis, Romy Sweda, Peter A. Noseworthy, Paul A. Friedman, Konstantinos C. Siontis, and Chirag J. Patel. "Development and validation pathways of artificial intelligence tools evaluated in randomised clinical trials." In: *BMJ Health & Care Informatics* 28.1 (Dec. 2021). ISSN: 2632-1009. DOI: 10.1136/bmjhci-2021-100466.

[249] Maria Filomena Santarelli, Dario Genovesi, Vincenzo Positano, Michele Scipioni, Giuseppe Vergaro, Brunella Favilli, Assuero Giorgetti, Michele Emdin, Luigi Landini, and Paolo Marzullo. "Deep-learning-based cardiac amyloidosis classification from early acquired pet images." In: *The International Journal of Cardiovascular Imaging* 37.7 (July 2021), pp. 2327–2335. DOI: 10.1007/s10554-021-02190-7.

[250] Alessandro Allegra, Giuseppe Mirabile, Alessandro Tonacci, Sara Genovese, Giovanni Pioggia, and Sebastiano Gangemi. "Machine Learning Approaches in Diagnosis, Prognosis and Treatment Selection of Cardiac Amyloidosis." In: *International Journal of Molecular Sciences* 24.6 (Mar. 2023), p. 5680. DOI: 10.3390/ijms24065680.

[251] Stephan Dobner et al. "Amyloid Transthyretin Cardiomyopathy in Elderly Patients With Aortic Stenosis Undergoing Transcatheter Aortic Valve Implantation." In: *Journal of the American Heart Association* 12.16 (Aug. 2023), e030271. DOI: 10.1161/JAHA.123.030271.

[252] Benedikt Bernhard et al. "Routine 4D Cardiac CT to Identify Concomitant Transthyretin Amyloid Cardiomyopathy in Older Adults with Severe Aortic Stenosis." In: *Radiology* 309.3 (Dec. 2023), e230425. ISSN: 0033-8419, 1527-1315.

[253] A. Pieter Kappetein et al. "Updated Standardized Endpoint Definitions for Transcatheter Aortic Valve Implantation." In: *Journal of the American College of Cardiology* 60.15 (Oct. 2012), pp. 1438–1454. DOI: 10.1016/j.jacc.2012.09.001.

[254] Philip Whybra et al. "The Image Biomarker Standardization Initiative: Standardized Convolutional Filters for Reproducible Radiomics and Enhanced Clinical Insights." In: *Radiology* 310.2 (Feb. 2024), e231319. DOI: 10.1148/radiol.231319.

[255] Alex Zwanenburg et al. "The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping." In: *Radiology* 295.2 (May 2020), pp. 328–338. DOI: 10.1148/radiol.2020191145.

[256] Joost J.M. Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. "Computational Radiomics System to Decode the Radiographic Phenotype." In: *Cancer Research* 77.21 (Nov. 2017), e104–e107. DOI: 10.1158/0008-5472.CAN-17-0339.

[257] Federico Caobelli, Nasir Gözlügöl, Adam Bakula, Axel Rominger, Robin Schepers, Stefan Stortecky, Lukas Hunziker Munsch, Stephan Dobner, and Christoph Gräni. "Prognostic Value of [^{99m} Tc]Tc-DPD Quantitative SPECT/CT in Patients with Suspected and Confirmed Amyloid Transthyretin–Related Cardiomyopathy and Preserved Left Ventricular Function." In: *Journal of Nuclear Medicine* 65.6 (June 2024), pp. 944–951. DOI: 10.2967/jnumed.123.266926.

[258] Christoph Gräni. "Early detection of subclinical cardiac amyloidosis: the importance of increasing physician awareness and routine imaging assessment." In: *The International Journal of Cardiovascular Imaging* 40.5 (May 2024), pp. 949–950. DOI: 10.1007/s10554-024-03148-1.

[259] Christoph Gräni. "Advancements in CT Tissue Characterization: Myocardial Insights in Aortic Stenosis and Amyloidosis." In: *Circulation: Cardiovascular Imaging* 17.5 (May 2024). DOI: 10.1161/CIRCIMAGING.124.016898.

[260] Jeffrey Zhang et al. "Fully Automated Echocardiogram Interpretation in Clinical Practice: Feasibility and Diagnostic Accuracy." In: *Circulation* 138.16 (Oct. 2018), pp. 1623–1635. DOI: 10.1161/CIRCULATIONAHA.118.034338.

[261] Shinichi Goto et al. "Artificial intelligence-enabled fully automated detection of cardiac amyloidosis using electrocardiograms and echocardiograms." In: *Nature Communications* 12.1 (May 2021), p. 2726. DOI: 10.1038/s41467-021-22877-8.

[262] Ahsan Huda, Adam Castaño, Anindita Niyogi, Jennifer Schumacher, Michelle Stewart, Marianna Bruno, Mo Hu, Faraz S. Ahmad, Rahul C. Deo, and Sanjiv J. Shah. "A machine learning model for identifying patients at risk for wild-type transthyretin amyloid cardiomyopathy." In: *Nature Communications* 12.1 (May 2021), p. 2725. DOI: 10.1038/s41467-021-22876-9.

[263] Hanna-Leena Halme, Toni Ihalainen, Olli Suomalainen, Antti Loimaala, Sorjo Mätzke, Valtteri Uusitalo, Outi Sipilä, and Eero Hippeläinen. "Convolutional neural networks for detection of

transthyretin amyloidosis in 2D scintigraphy images." In: *EJNMMI Research* 12.1 (Dec. 2022), p. 27. DOI: 10.1186/s13550-022-00897-9.

[264] Marc-Antoine Delbarre et al. "Deep Learning on Bone Scintigraphy to Detect Abnormal Cardiac Uptake at Risk of Cardiac Amyloidosis." In: *JACC: Cardiovascular Imaging* 16.8 (Aug. 2023), pp. 1085–1095. DOI: 10.1016/j.jcmg.2023.01.014.

[265] Nicola Martini, Alberto Aimo, Andrea Barison, Daniele Della Latta, Giuseppe Vergaro, Giovanni Donato Aquaro, Andrea Ripoli, Michele Emdin, and Dante Chiappino. "Deep learning to diagnose cardiac amyloidosis from cardiovascular magnetic resonance." In: *Journal of Cardiovascular Magnetic Resonance* 22.1 (Jan. 2020), p. 84. DOI: 10.1186/s12968-020-00690-4.

[266] Asan Agibetov et al. "Convolutional Neural Networks for Fully Automated Diagnosis of Cardiac Amyloidosis by Cardiac Magnetic Resonance Imaging." In: *Journal of Personalized Medicine* 11.12 (Dec. 2021), p. 1268. DOI: 10.3390/jpm11121268.

[267] Francesca Lo Iacono, Riccardo Maragna, Gianluca Pontone, and Valentina D. A. Corino. "A robust radiomic-based machine learning approach to detect cardiac amyloidosis using cardiac computed tomography." In: *Frontiers in Radiology* 3 (June 2023), p. 1193046. DOI: 10.3389/fradi.2023.1193046.

[268] Francesca Lo Iacono et al. "Identification of subclinical cardiac amyloidosis in aortic stenosis patients undergoing transaortic valve replacement using radiomic analysis of computed tomography myocardial texture." In: *Journal of Cardiovascular Computed Tomography* 17.4 (July 2023), pp. 286–288. DOI: 10.1016/j.jcct.2023.04.002.

[269] Esther González-López et al. "Wild-type transthyretin amyloidosis as a cause of heart failure with preserved ejection fraction." In: *European Heart Journal* 36.38 (Oct. 2015), pp. 2585–2594. DOI: 10.1093/eurheartj/ehv338.

[270] Steven A. Muller et al. "Absence of an increased wall thickness does not rule out cardiac amyloidosis." In: *Amyloid* 31.3 (July 2024), pp. 244–246. DOI: 10.1080/13506129.2024.2348681.

[271] Achala Donuru et al. "Photon-counting detector CT allows significant reduction in radiation dose while maintaining image quality and noise on non-contrast chest CT." In: *European Journal of Radiology Open* 11 (Dec. 2023), p. 100538. DOI: 10.1016/j.ejro.2023.100538.

[272] Catherine M. Otto et al. "2020 ACC/AHA Guideline for the Management of Patients With Valvular Heart Disease: Executive Summary: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clini-

cal Practice Guidelines." In: *Circulation* 143.5 (Feb. 2021). DOI: 10.1161/CIR.0000000000000932.

[273] Clemens P Spielvogel et al. "Diagnosis and prognosis of abnormal cardiac scintigraphy uptake suggestive of cardiac amyloidosis using artificial intelligence: a retrospective, international, multicentre, cross-tracer development and validation study." In: *The Lancet Digital Health* 6.4 (Apr. 2024), e251–e260. DOI: 10.1016/S2589-7500(23)00265-0.

[274] Manesh R. Patel, Suresh Balu, and Michael J. Pencina. "Translating AI for the Clinician." In: *JAMA* (Oct. 2024). DOI: 10.1001/jama.2024.21772.

[275] Chava L Ramspek, Kitty J Jager, Friedo W Dekker, Carmine Zoccali, and Merel van Diepen. "External validation of prognostic models: what, why, how, when and where?" In: *Clinical Kidney Journal* 14.1 (Jan. 2021), pp. 49–58. DOI: 10.1093/ckj/sfaa188.

[276] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. "Key challenges for delivering clinical impact with artificial intelligence." In: *BMC Medicine* 17.1 (Oct. 2019), p. 195. DOI: 10.1186/s12916-019-1426-2.

[277] Joseph Futoma, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. "The myth of generalisability in clinical research and machine learning in health care." English. In: *The Lancet Digital Health* 2.9 (Sept. 2020), e489–e492. DOI: 10.1016/S2589-7500(20)30186-2.

[278] Ana Sofia Figueiredo. "Data Sharing: Convert Challenges into Opportunities." In: *Frontiers in Public Health* 5 (Dec. 2017). DOI: 10.3389/fpubh.2017.00327.

[279] Bradley Malin, Kenneth Goodman, and Section Editors for the IMIA Yearbook Special Section. "Between Access and Privacy: Challenges in Sharing Health Data." In: *Yearbook of Medical Informatics* 27.01 (Aug. 2018), pp. 055–059. DOI: 10.1055/s-0038-1641216.

[280] *UK Biobank - UK Biobank*. en-gb. Oct. 2024. URL: https://www.ukbiobank.ac.uk (visited on 10/17/2024).

[281] Shinjini Kundu. "AI in medicine must be explainable." In: *Nature Medicine* 27.8 (Aug. 2021), pp. 1328–1328. ISSN: 1546-170X. DOI: 10.1038/s41591-021-01461-z.

[282] Sandeep Reddy. "Explainability and artificial intelligence in medicine." In: *The Lancet Digital Health* 4.4 (Apr. 2022), e214–e215. DOI: 10.1016/S2589-7500(22)00029-2.

[283] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I. Madai, and the Precise4Q consortium. "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective." In: *BMC Medical Informatics and Decision Making* 20.1 (Nov. 2020), p. 310. DOI: 10.1186/s12911-020-01332-6.

[284] Liam G. McCoy, Connor T. A. Brenna, Stacy S. Chen, Karina Vold, and Sunit Das. "Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based." In: *Journal of Clinical Epidemiology* 142 (Feb. 2022), pp. 252–257. DOI: 10.1016/j.jclinepi.2021.11.001.

[285] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L. Beam. "The false hope of current approaches to explainable artificial intelligence in health care." In: *The Lancet Digital Health* 3.11 (Nov. 2021), e745–e750. DOI: 10.1016/S2589-7500(21)00208-9.

[286] Nicola Rieke et al. "The future of digital health with federated learning." In: *npj Digital Medicine* 3.1 (Sept. 2020), pp. 1–7. ISSN: 2398-6352. DOI: 10.1038/s41746-020-00323-1.

[287] Theodora S. Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch. Paschalidis, and Wei Shi. "Federated learning of predictive models from federated Electronic Health Records." In: *International Journal of Medical Informatics* 112 (Apr. 2018), pp. 59–67. DOI: 10.1016/j.ijmedinf.2018.01.007.

[288] Akis Linardos, Kaisar Kushibar, Sean Walsh, Polyxeni Gkontra, and Karim Lekadir. "Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease." In: *Scientific Reports* 12.1 (Mar. 2022), p. 3551. DOI: 10.1038/s41598-022-07186-4.

[289] Rohan Khera and Jenna Wiens. "Summertime for Cardiovascular AI." In: *Circulation: Cardiovascular Quality and Outcomes* 17.3 (Mar. 2024), e010404. DOI: 10.1161/CIRCOUTCOMES.123.010404.

[290] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. "Foundation models for generalist medical artificial intelligence." In: *Nature* 616.7956 (Apr. 2023), pp. 259–265. DOI: 10.1038/s41586-023-05881-4.