Carlos Colmenero Gómez Cambronero

Fair cardiovascular disease diagnosis and prognosis through machine learning and fractal-based features

MASTER'S THESIS

Supervised by Dr. Polyxeni Gkontra

Master's Degree in Biomedical Data Science

Tarragona, 2024

I would like to express my deep gratitude to my supervisor, Xenia, for her guidance over the past several months. Her patience and kindness have been a constant source of support, and her insightful advice has been immensely valuable to me. Thank you, Xenia, for introducing me to the fascinating world of research.

I am also deeply thankful to my family and friends for their unwavering support throughout my academic endeavors, providing encouragement and understanding at every step of the way.

My heartfelt thanks go to my colleagues, Lucía, Eric, and Judith, for showing me humanity in a virtual world. This master's degree journey would not have been the same without you.

And, finally, I am grateful to science for teaching me that we indeed stand on the shoulders of giants, a perpetual lesson in humility.

Thank you all.

Abstract

Given the high number of deaths caused by cardiovascular diseases, innovative methods are essential to mitigate their detrimental effects. Fractal analysis holds promise in this area by providing a detailed representation of complex patterns, as the ones representative of cardiovascular conditions. In this thesis, the fairness and predictive performance of ML models utilizing fractal-based features derived from CMR was evaluated for the diagnosis and prognosis of cardiovascular diseases in a sample of 31,931 subjects of the UK-Biobank. The performance of the fractal models was compared to widely-used features like CMR indices and radiomics. Several machine learning models were used, including gradient boosting, and bagging models. The ML pipeline involved a two-level 3-fold cross-validation (CV), with the inner CV for hyperparameter tuning (Bayesian-based) and the outer CV for testing. Fairness was assessed with the average odds difference (AOD) as the main metric, and mitigation was applied through the exponentiated gradient technique. The best fractal models were inspected through SHAP feature importance. No statistical significant differences were found between the best fractal and radiomics models for any disease, and fractal features were in the feature set of the best-performing models for most diseases. After fairness mitigation, fractal models were superior to radiomics models in terms of AOD. Considering the equivalent predictive performance, the lower bias if mitigation is applied, and the smaller number of features, fractals are proposed as an alternative to radiomics for future research. Myocardium complexity and heterogeneity were the most important fractal features, and right ventricle fractal features were revelant for two diseases in prognosis, suggesting that the right ventricle should not be as ignored as it has historically been for cardiovascular diseases.

Keywords: Machine learning, fractal-based features, cardiovascular diseases, diagnosis, prognosis, fairness

Dr. Polyxeni Gkontra certifies that the student Carlos Colmenero Gómez Cambronero has elaborated the work under her direction and she authorizes the presentation of this Master's Thesis for its evaluation.

Advisor's signature:

GKONTRA --- POLYXENI - Y3425665J

Digitally signed by GKONTRA --- POLYXENI - Y3425665J
Date: 2024.06.07 06:43:12 +02'00'

Table of Contents

List of Figures

List of Tables

Abbreviations and acronyms

AOD Average Odds Difference

BRF Balanced Random Forest

CMR Cardiac Magnetic Resonance

CV Cross-Validation

GLCM Gray-Level Co-occurrence Matrix

GLDM Gray Level Dependence Matrix

GLRLM Gray-Level Run-Length Matrix

GLSZM Gray-Level Size Zone Matrix

HF Heart Failure

LGBM Light Gradient Boosting Machine

MI Myocardial Infarction

ML Machine Learning

NGTDM Neighborhood Gray-Tone Difference Matrix

RF Random Forest

SVM Support Vector Machine

VHD Valvular Heart Disease

XGB Extreme Gradient Boosting

Chapter 1

Motivation and Background

1.1 Motivation

Although the mortality rate from cardiovascular diseases has decreased in recent years, they still represent the leading cause of death worldwide, contributing to one third of all deaths [1]. In 2019, it is estimated that 18.6 million deaths were caused by cardiovascular diseases, of which 9.14 million were from ischemic heart disease [2].

The methods used to identify cardiovascular pathologies have undergone important changes over time. Initially, evaluations were performed by physical examinations that included palpation of the pulse and auscultation of heart sounds [3] [4]. Then, electrocardiography allowed to record the electrical activity of the heart, so that diseases such as arrhythmias could be more easily detected. However, these techniques had a very limited view, since they did not precisely capture the structure and internal functioning of the cardiovascular system. This problem was addressed with the adoption of advanced imaging methods like computed tomography (CT) and cardiac magnetic resonance (CMR) [5] [6].

The introduction of CMR imaging in the 1980s marked a paradigm shift, as it enabled the creation of high-resolution images in a non-invasive manner and without the ionizing radiation associated with CT and nuclear cardiology (PET and SPECT), reducing the negative health impact from radiation exposure [6].

CMR, like other magnetic resonance imaging techniques, involves applying a magnetic field to align the protons of the hydrogen nuclei in water molecules, and emitting radio frequency waves to perturb the alignment of the protons with respect to the magnetic field. When the radio waves are turned off, the hydrogen nuclei return to their original alignment, releasing energy in the form of radio frequency that is captured and used to generate images of the target area [7].

Conventional CMR indices, which are quantitative measurements obtained from CMR images, have been utilized for years to diagnose and prognosticate cardiovascular pathologies. These indices provide relevant information about the heart's shape, function, and tissue characteristics [8] [9]. They typically include volumetric indices (such as ejection fraction, and end-diastolic or end-systolic volumes), myocardial masses (like left ventricular mass), as well as functional indices (such as stroke volume and cardiac output).

However, the complex nature of cardiovascular diseases requires more sophisticated tools to identify detailed patterns that often ignore traditional diagnostic methods. In recent years, radiomics has emerged as a compelling area of research for this purpose. It involves the segmentation of the region of interest (ROI), followed by the extraction of quantitative features from the ROI, and the later analysis of the extracted features [10]. Radiomics can help in the detection of complex conditions like cardiovascular diseases by providing relevant insights about tissue characteristics and lesion signatures that are challenging for the human eye to detect [11] [12].

An alternative method to elucidate the intricate patterns present in tissues utilizes fractals, which are geometric figures that can be divided into parts, with each part being a reduced replica or similar version of the whole structure [13]. These patterns are prevalent in nature, such as the branching of tree limbs or the spirals of a seashell. In the human cardiovascular system, fractal structures are observed as well, including the branching patterns of blood vessels, the structure of the muscle fibers lining the inner walls of the heart's ventricles, and the arrangement of collagen and fibrosis within the diseased myocardial tissue [14]. Fractal analysis has generated promising results in brain imaging regarding neurodegenerative diseases and brain tumor detection [15] [16] [17] [18]. However, its potential in cardiovascular research and CMR imaging has been much less explored [14].

Machine Learning (ML) is a subfield of artificial intelligence that focuses on enabling machines to learn and take decisions based on data [19]. In the biomedical domain, ML has been employed for applications such as drug discovery and genomics, paving the way towards higher quality and increasingly personalized medicine [20]. One of its most widely used applications is disease diagnosis and prediction. Several ML models have been able to accurately predict the diagnosis of different several using different types of data like blood test information or medical images [21], as well as their occurrence in the future, which poses a much bigger challenge [22].

Given the high number of deaths caused by cardiovascular diseases, innovative methods are essential to mitigate their detrimental effects. Fractal analysis holds promise in this area by providing a detailed representation of the complex patterns representative of cardiovascular diseases. Combining fractal features with machine learning techniques could potentially improve the precision of diagnosing and prognosticating cardiovascular conditions. Also, it would be attractive to explore their possible incremental value in combination with the traditional CMR indices that have been used for years. Furthermore, it would be interesting to evaluate whether fractal-based features produce similar results to those obtained with radiomics when combined with machine learning, which have already proven to be useful in cardiovascular research [23] [24] [25], as well as their potential added value when including them together. To ensure a complete comparison, it would be relevant to evaluate not only the predictive performance but also the fairness of the models, a research field that has gained recent attention [26] [27].

1.2 Background

1.2.1 Image-based features

In this section, the utilized feature set will be described, including conventional fractal-based features, CMR indices, and radiomics-based features.

Fractal Features

Fractals are patterns that keep their general shape regardless of the scale [28]. Figure 1.1 illustrates one of the most famous fractals: the *Koch curve*. It reflects how each step increases the complexity of the shape without changing its overall geometry.

Figure 1.1: Koch curve with increasing number of iterations.

Fractal dimensions and Lacunarity

Fractal dimensions are measurements that quantify how completely a fractal fills space in increasingly smaller scales, which reflects the the degree of detail of the fractal pattern. They can be used to capture the complexity of intricate structures such as a cardiac volume [29].

Lacunarity is a measure of how the texture or gap size distribution of a fractal pattern changes across different scales, which reflects its degree of heterogeneity. It can provide insights into the spatial distribution and density of the tissue, which might be indicative of changes associated with some cardiovascular diseases [30].

Conventional CMR indices

Conventional CMR indices provide quantitative assessments of cardiac structure and function through several measurements [31]. Some of the most commonly used CMR indices include:

- **End-diastolic volume** (LV and RV). It measures the volume of blood within the ventricles at the completion of the diastolic phase, reflecting the maximum volume of blood the ventricles hold before contraction.
- **End-systolic volume** (LV and RV). It quantifies the volume of blood remaining in the ventricles at the end of the systolic phase, indicating the volume left after contraction.
- **Stroke volume** (LV and RV). It is the volume of blood ejected from the ventricles with each heartbeat
- **Ejection fraction** (LV and RV). It is the ratio of stroke volume compared to end-diastolic volume.
- **LV mass**. It is the mass of the left ventricular myocardium.

Radiomics Features

Radiomics allows for the extraction of a large number of quantitative features from medical images that are hard to identify by the naked eye, describing characteristics such as the shape, texture, and intensity of image regions [10] [32]. The main classification of the types of radiomics features, as well as an explanation of each of them, is detailed below:

- **First-order statistics**. They are based on the distribution of individual voxel intensities, and include basic statistics such as mean, median, variance, minimum and maximum values, skewness, and kurtosis.
- **Second-order statistics**. They are also known as *textural features*, and are derived from the relationships between nearby voxels. These features help quantify how voxel intensities are spatially organized, which in turn provides insights about the heterogeneity of the lesion. This category includes:

  - **Gray-Level Cooccurrence Matrix (GLCM)**: Analyzes voxel intensity pairs that are separated by a specific distance.
  - **Gray-Level Run-Length Matrix (GLRLM)**: Measures the length of contiguous *runs*, which refer to the number of voxels with similar gray levels.
  - **Gray-Level Size Zone Matrix (GLSZM)**: Evaluates the *zones* of an image, which are connected voxels with similar gray levels.
  - **Neighborhood Gray-Tone Difference Matrix (NGTDM)**: Assesses the discrepancy between a gray level and the surrounding mean gray level.
  - **Gray Level Dependence Matrix (GLDM)**: Measures the *dependence* of gray levels, by counting the number of neighbors whose intensity difference from a central voxel is less than a specified threshold.

- **Shape-based features**. They describe the geometric properties of the region of interest, such as area, volume, diameter, elongation, and flatness.

1.2.2 Machine Learning Models

Support Vector Machine

Linear Support Vector Machine

Support Vector Machine (SVM) is a machine learning model that tries to find the hyperplane that better separates the classes in the feature space [33]. In SVM, the predictions are obtained from the decision function, which in the linear case consists of the dot product of the weight vector and the feature vector, as well as a bias term. The weight vector is perpendicular to the hyperplane and determines its orientation, whereas the bias term considers shifts with respect to the origin.

$$\hat{y}_i = \mathbf{w} \cdot \mathbf{x}_i + b, \tag{1.1}$$

- $\mathbf{w}$ is the weight vector.
- $\mathbf{x_i}$ is the feature vector of the ith observation.
- b is the bias term.

SVM tries not only to separate the classes but also to maximize the distance between the hyperplane and the nearest points of any class, which is known as the **margin**. In this way, there is a safety distance that could help in distinguishing the classes of unseen observations that are more difficult to classify. The margin is computed based on the magnitude of the weight vector.

$$M = \frac{2}{\|\mathbf{w}\|} \tag{1.2}$$

where,

- $\|\mathbf{w}\|$ is the magnitude of the weight vector.

The **support vectors**, from which the SVM name comes from, are the data points that lie on upper or lower boundaries of the margin, where the decision value is ± 1. A visual representation of the hyperplane, the margin, and the support vectors is illustrated in figure 1.2.

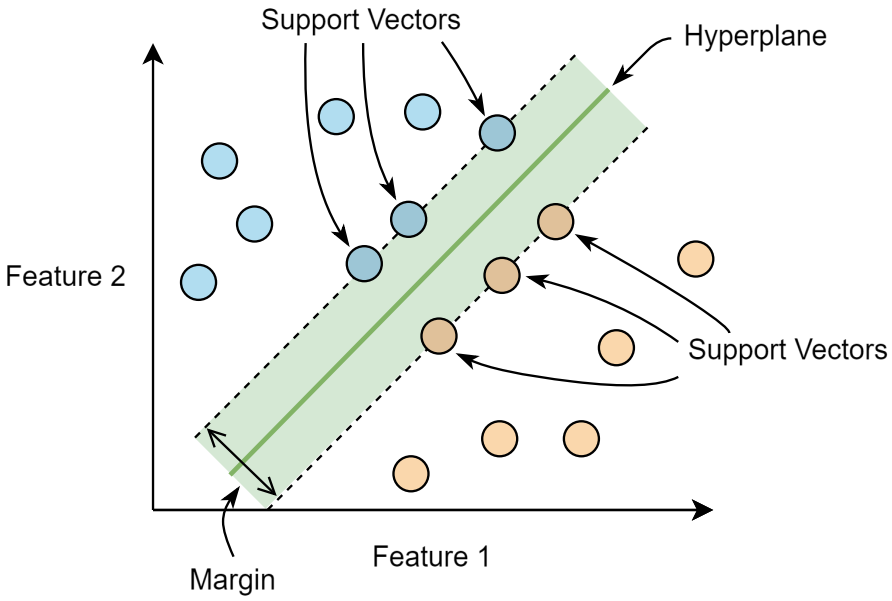


Figure 1.2: Main elements of SVM, including the hyperplane (represented with a green line), the margin (drawn with a double headed arrow), and the support vectors (shown in dark blue and brown).

As the loss function to optimize, SVM employs the **hinge loss**, which is designed to penalize both incorrect predictions and predictions inside the margin.

$$L(y_i, \hat{y}_i) = \max(0, 1 - y_i(\mathbf{w} \cdot \mathbf{x}_i + b)), \tag{1.3}$$

where,

- $\hat{y}_i$ is the prediction of the model.
- y_i is the real label.
- $\mathbf{w}$ is the weight vector.
- $\mathbf{x_i}$ is the feature vector in the ith observation.

The optimal separation is found by minimizing an objective function that includes a loss function and a regularization term. By scaling the loss function, the hyperparameter **C** penalizes incorrect classifications, and can be used to control the trade-off between prioritizing the loss or the margin. A higher C value leads to fewer classification errors but a smaller margin, while a lower C leads to more classification errors but a larger margin. The maximization of the margin is achieved by including in the objective function (that is minimized) a term based on the weight vector, which is inversely proportional to the margin.

$$\text{Obj}(\mathbf{w}, b) = C \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \frac{1}{2}|\mathbf{w}|^2, \tag{1.4}$$

where,

- $\hat{y}_i$ is the prediction of the model.
- y_i is the real label.
- $\mathbf{w}$ is the weight vector of the hyperplane.
- C is the regularization parameter.
- n is the number of observations in the dataset.

Kernel SVM

Sometimes the data cannot be linearly separated in the original feature space. In these situations, a kernel function help SVMs to map features into a higher-dimensional space where the data can be linearly separated, which is usually known as the *kernel trick* [34]. The **kernel function** measures similarities between data points, and uses them to compute the dot products in a higher-dimensional space without explicitly transforming the feature vector into that space. The kernel function measures similarities between data points, which represent the dot product in a higher-dimensional space, without explicitly transforming the feature vector into that space. Lagrange multipliers are usually used to ensure that the margin constraints are met, which make the optimization more efficient. In kernel SVM, the decision function is computed as:

$$\hat{y}_i = \sum_{j=1}^{n} \alpha_j y_j K(\mathbf{x}_j, \mathbf{x}_i) + b, \tag{1.5}$$

where:

- $\hat{y}_i$ determines is the predicted class for the i-th observation. If $\hat{y}_i > 0$, the prediction is 1, whereas it is -1 if $\hat{y}_i < 0$.
- α_j are the Lagrange multipliers.
- $K(\mathbf{x}_j, \mathbf{x}_i)$ is the kernel function.
- b is the bias term.

Objective function for kernel SVM is based on finding the optimal set of Lagrange multipliers that maximize the margin:

$$\text{Obj} = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j), \tag{1.6}$$

where:

- $\boldsymbol{\alpha}$ are the Lagrange multipliers to optimize.
- y_i and y_j are the real labels of the ith and jth observations, respectively.
- $K(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel function.

Introduction to Bagging and Boosting

A **decision tree** is a machine learning model whose main functioning consists of evaluating if a condition is fulfilled [35]. If the condition is met, the model continues with one branch, following with the other branch otherwise. The process starts at the *root node*, which represents the whole dataset, and the data is divided based on the feature with the highest gain. The *gain* is the difference in the uncertainty of the model predictions before and after splitting using a certain *criterion* (for example, Gini impurity or entropy) [36]. As the tree grows, the dataset is divided into smaller and smaller subsets, until reaching the terminal nodes (also known as *leaves*), which represent the final decisions.

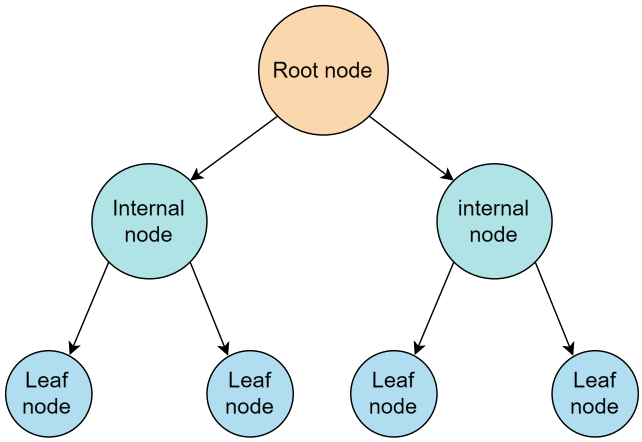


Figure 1.3: Decision tree scheme in which the hierarchy of its nodes can be observed, from the root node to the leaf nodes.

There are models, called **ensembles**, that combine the predictions of several base learners to decrease the variance and the bias compared to a single model [37]. One example of them are tree-based models, which are ensembles that uses decision trees as their base learners [38].

There are different types of ensembles, including bagging and boosting. **Bagging** stands for bootstrap aggregating, involves creating several subsets of the original dataset using bootstrapping, in which subsets are sampled with replacement (i.e. allowing the same observation to appear more than once) [39]. The base models are trained independently on the sampled subsets, so that each model has a different view of the data. After all models are trained, their predictions are combined to create the final result, which consists of majority voting in classification.

Gradient-based models employ **boosting**, a method where new models are added in a sequential manner to improve the predictive performance by minimizing the errors from previous models or giving more weight to observations that were previously incorrectly classified [40] [41]. Once all models are trained, they are weighted according to how they contributed to reduce the errors and used to generate the final prediction.

Figures 1.4 and 1.5 show the main difference between both ensemble methods, where bagging follows a parallel approach in which the decision trees are trained independently, whereas boosting utilizes a sequential approach where the decision trees consider information from the previous ones.

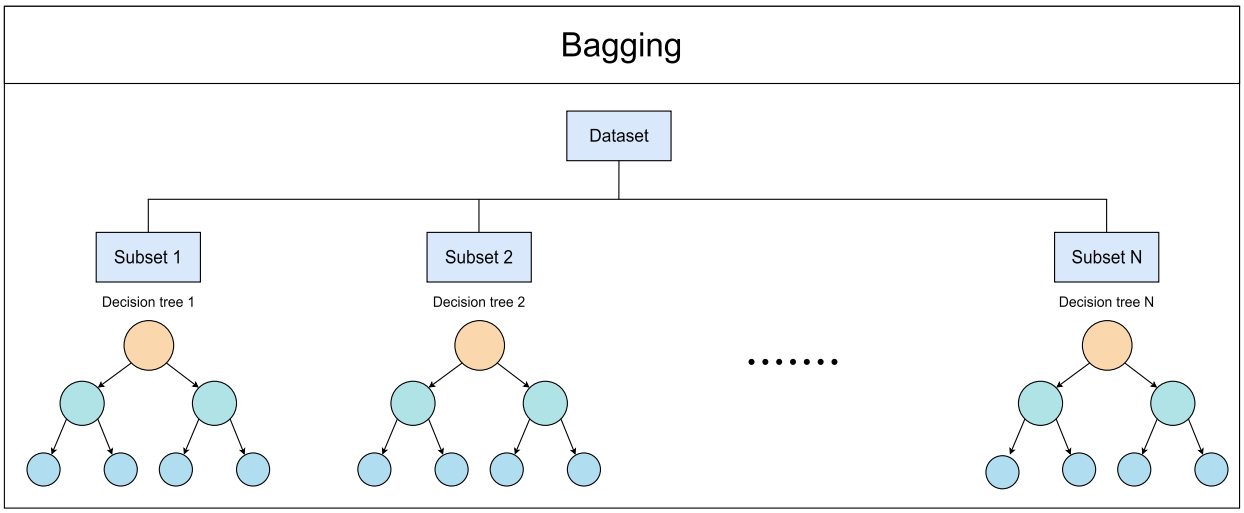


Figure 1.4: Bootstrap aggregating (bagging) scheme, which follows a parallel approach.

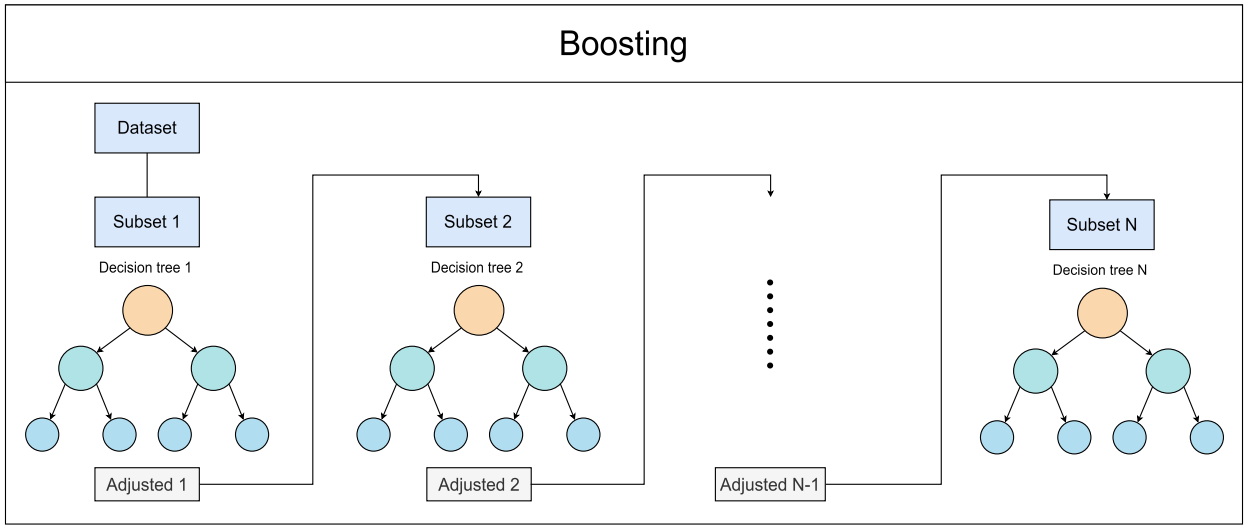


Figure 1.5: Gradient boosting scheme, which follows a sequential approach.

In this thesis, both bagging (random forest and balanced random forest) and gradient-based models (XGBoost, LightGBM and AdaBoost) were utilized.

Bagging and Boosting Models

XGBoost

XGBoost [42], which stands for eXtreme Gradient Boosting, tries to minimize an objective function that is composed of two parts: the loss function $L(y, \hat{y})$, which is calculated for all observations, and the regularization term $\Omega(f_k)$, which is summed for all trees.

$$\text{Obj}(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k) \tag{1.7}$$

The used loss function was the **log loss**, which is suitable for binary classification tasks, like the one we are dealing with (diseased versus not diseased). It quantifies the uncertainty of the model predictions by measuring how much the predicted probabilities differ from the real labels. The *log loss* function for binary classification is defined as follows:

$$L(y, \hat{y}) = -\frac{1}{n} \sum_{i=1}^{n} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{1.8}$$

where:

- y_i is the actual label of the instance, that can be 0 or 1.
- $\hat{y}_i$ is the predicted probability that the observation i belongs to the positive class (label 1).
- The sum is computed over all n observations in the dataset.

The **regularization term** $\Omega(f_k)$ helps to control overfitting by penalizing the complexity of the model in terms of the number of leaves. Additionally, a L2 regularization term is included, which penalizes the sum of the squares of the leaf weights, promoting low weight values. Optionally, an L1 regularization term can be included, which minimizes the absolute value of the leaf weights and can act as pruning by making some weights to be zero. The regularization term for a given tree f_k is defined as follows:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 + \alpha \sum_{j=1}^{T} |w_j| \tag{1.9}$$

where:

- T is the total number of leaves in the tree,
- w_j are the weights of each leaf,
- γ is a regularization parameter that controls the model complexity by penalizing the number of leaves,
- λ is a regularization parameter that controls the magnitude of the weights (L2 regularization),
- α is a regularization parameter that allows that promotes that some leaf weights are zero (L1 regularization).

XGBoost employs the gradient and hessian of the loss function to decide the best way to split the data. The gradient and the hessian for the log loss are calculated as:

$$g_i = \frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}_i} = \hat{y}_i - y_i \tag{1.10}$$

where:

- g_i is the gradient of the loss for the i-th observation.

$$h_i = \frac{\partial^2 L(y_i, \hat{y}_i)}{\partial \hat{y}_i^2} = \hat{y}_i(1 - \hat{y}_i) \tag{1.11}$$

where:

- h_i is the Hessian (second derivative) of the loss for the i-th observation.

Then, the **gain** can be computed. The best split is the one with which the higher gain is obtained.

$$\text{Gain} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma, \tag{1.12}$$

where:

- I_L and I_R are indices of observations to the left and right children, respectively.

Like other boosting models, the trees of the XGBoost are constructed additively, so that each tree is fit on the residual errors made by the previous trees. The prediction at iteration t is:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \tag{1.13}$$

where:

- $\hat{y}_i^{(t-1)}$ is the prediction from the previous iteration,
- f_t is the new tree,
- η is the learning rate.

LightGBM

The main charasteristic of LightGBM is its balance between efficiency and predictive performance [43]. In general terms, it is similar to XGBoost. For instance, they use an alike objective function, including the loss function and the regularization term.

However, unlike traditional gradient boosting models (like XGBoost) that use all observations for computing the gradients, LightGBM uses only the most challenging instances (those with higher loss). It employs the **Gradient-based One Side Sampling** approach, which consists of considering all the observations with large gradients and randomly sampling some with low gradients. In this way, the focus is on the data points that are harder for the model to classify, and less importance is given to the ones it classifies more accurately.

LightGBM also helps to handle large datasets by reducing the number of features. It does this through the **Exclusive Feature Bundling** method, in which features that are mutually exclusive (i.e., features that are unlikely to be non-zero at the same time) are grouped together.

In XGBoost, trees are built by growing all the nodes in each level, until a certain depth is met (level-wise). In contrast, the decision trees in LightGBM are grown by iteratively adding the best leave (**leaf-wise**), as illustrated in figure 1.7. Compared to growing the trees in an uniform way, this method is more memory efficient and could potentially lead to a less overfitting, as unnecessary leaves are not included.

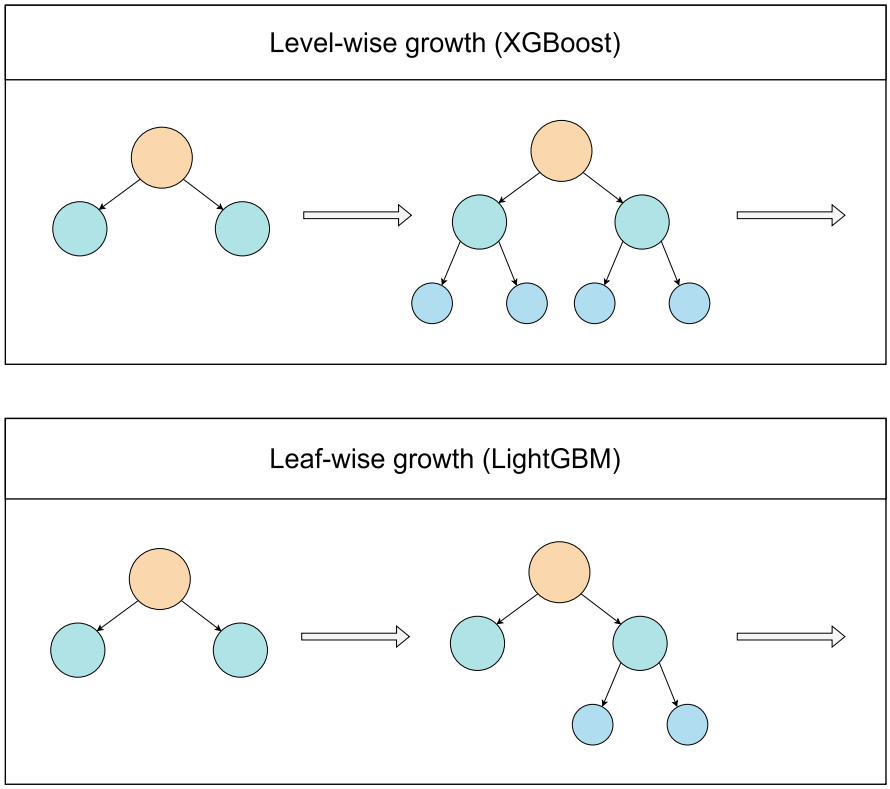


Figure 1.6: Level-wise (top) and leaf-wise (bottom) decision tree growth approaches.

In addition, LightGBM is more efficient in its management of continuous features, as it discretizes them into bins, which speed up the calculation of the best splits (since the number of possible splits is decreased).

AdaBoost

Adaptive Boosting (AdaBoost), instead of being designed to minimize loss functions with regularization terms like XGBoost and LightGBM, is an iterative algorithm that focuses on adjusting the weights of the observations based on the performance of the previous iterations [44]. AdaBoost uses two different types of weights: for base estimators, and for the observations. The weights of the base estimators measure their influence in the final model, and are computed with the following formula, which is based on the error rate:

$$\alpha_t = \frac{1}{2} \log \left(\frac{1 - \epsilon_t}{\epsilon_t} \right), \tag{1.14}$$

where:

- ϵ_t is the error rate of the t-th base estimator, calculated as the sum of the weights of the observations that it incorrectly classified, divided by the total sum of the weights of all observations in the current iteration.

After the weight of the base estimator is computed, the weights of the observations are updated. AdaBoost adapts quickly to observations that were previously incorrectly classified by assigning them exponentially higher weights, while the weights of correctly classified observations are decreased.

$$D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t}, \tag{1.15}$$

where:

- $D_t(i)$ is the weight of the i-th observation at iteration t,
- α_t is the weight of the t-th base estimator,
- y_i is the real class label,
- $h_t(x_i)$ is the prediction of the base estimator,
- Z_t is a normalization factor that makes D_{t+1} to be a probability distribution.

The final prediction of the model is made based on a weighted majority vote of its base estimators. As AdaBoost is designed for the output of its base estimators to be either -1 or 1, the final output of the ensemble is decided based on the sign of the weighted majority voting.

$$f(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right), \tag{1.16}$$

where:

- $h_t(x)$ is the prediction of the t-th base estimator, which can be -1 or 1.
- α_t is the weight of the t-th base estimator.
- T is the total number of base estimators.

Although the traditional output of AdaBoost for binary classification is -1 or 1, in practical applications these values are usually mapped to 0 and 1 to be consistent with the output of other binary classifiers.

Random Forest

Random Forest is an ensemble model that provides each individual tree a different view of the data [45]. This is achieved through sampling with replacement (**bootstrapping**), in which observations are randomly selected, allowing the same instance to appear more than once. Each tree uses a different bootstrap sample. In addition, a subset of features is randomly selected for each base learner.

Gini impurity was used as criterion to decide the splits. As shown in equation 1.17, it measures the probability that an observation is not correctly classified if it was randomly labeled according to the class distribution of the target variable. For binary classification, it ranges from 0 to 0.5, where 0 represents perfect purity (all observations belong to the same class), and 0.5 the maximum impurity (the classes are equally distributed). Random forest tries to maximize the Gini impurity decrease (which is similar to maximize the information gain), trying to separate the data so that the subsets are more homogeneous than the set before the split.

$$\text{Gini} = 1 - \sum_{i=1}^{J} p_i^2, \tag{1.17}$$

where:

- p_i is the probability of an observation being classified as class i.
- J is the number of classes.

In classification, the final prediction is generated through a majority voting of all the trees.

$$\text{Final Class} = \arg\max_{c \in \{1,2\}} \sum_{i=1}^{N} I(y_i = c), \tag{1.18}$$

where:

- y_i is the prediction of the i-th tree,
- N is the total number of trees,
- I is 1 if $y_i = c$ and 0 otherwise.

Balanced Random Forest

Balanced Random Forest (BRF) is a variation of random forest designed for unbalanced scenarios whose main difference is the way it performs sampling [46]. Regular random forest conducts random sampling, so in unbalanced datasets, the majority classes are more represented. In contrast, balanced random forest ensures a balanced frequency of the classes in the training data of each tree, which could prevent that the model is biased to the majority class. BRF can be used with several sampling strategies, including oversampling or undersampling [47].

1.2.3 Bayesian Optimization

There are several techniques to optimize the hyperparameters of the models, with some of the most widely used being grid search and Bayesian optimization. As opposed to grid search, which evaluates all possible values and always finds the optimal ones (although being time-consuming), Bayesian optimization finds nearly optimal values in an efficient manner by making informed decisions about the next values to evaluate [48].

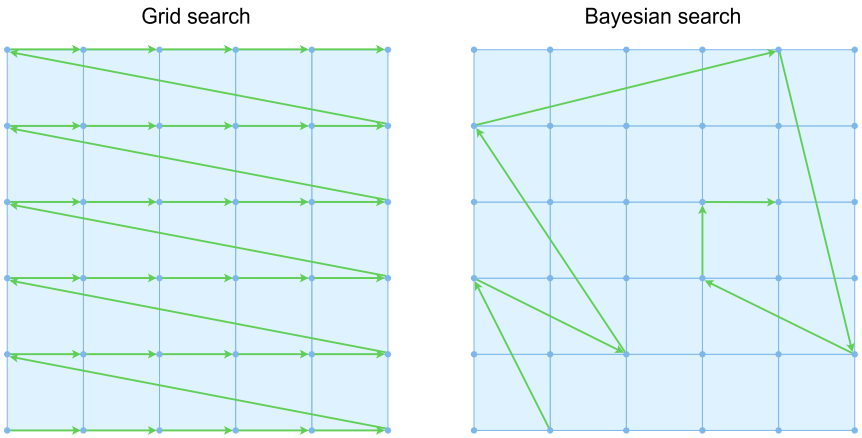


Figure 1.7: Grid search (left) in which all values are tried, and Bayesian search (right) in which only selected values are tried based on the previous knowledge.

In Bayesian optimization, the objective function is usually modeled using a Gaussian Process (GP), which starts with a prior distribution based on assumptions about the behaviour of the objective function, and updates to a posterior distribution when new results are obtained [49]. The posterior distribution is used to make predictions as well as compute their variability. The *acquisition function* chooses the next hyperparameter value by considering both the exploration of low certainty regions and the exploitation on areas where the GP model predicts better performance. When new data is available, the posterior distribution is updated and the acquisition function selects the new value that will be evaluated. This process is repeated for a specific number of iterations or until a stopping criterion is reached.

In this thesis, the used optimization method was **Tree-structured Parzen Estimator** (TPE), which differs from traditional Gaussian Processes although having a Bayesian basis [50]. Its name comes from its use of kernel density estimators (also known as Parzen estimators) to model the distribution of hyperparameters, which allows a flexible non-parametric approximation of value distributions with empirical data, as well as its hierarchical tree structure of the search space. TPE employs two different density functions, $l(x)$ for better values of the objective functions (higher predictive performance), and $g(x)$ for worse values (lower performance). Those density functions are used to sample the hyperparameters based on the ratio $\frac{l(x)}{g(x)}$, where a higher ratio means that the hyperparameters are likely to get better performance. Sampling is performed in areas that have the potential for better performance (with higher ratio) but which have not been as sampled as other regions. One advantage of TPE over other methods like GP is that it can deal with both discrete and categorical variables, since it has not smoothness assumptions.

Optuna is a machine learning optimization framework that implements different methods to determine the optimal parameters, such as median stopping, successive halving, and Bayesian optimization [51]. The general functioning of Optuna consists of maximizing an objective function which includes fitting the model and computing and returning the predictive performance on validation (or minimizing the validation loss).

1.2.4 Fairness

Fairness Assessment

Traditionally, machine learning models were evaluated exclusively in terms of predictive performance. However, in recent years, importance has begun to be given not only to predictive power, but also to how fair the models are [52]. This is especially relevant in models whose predictions involve people, such as predictive models for the diagnosis or prognosis of diseases [53]. Assessing the fairness of the model could prevent bias or discrimination in features like age or sex, which are usually referred to as protected or *sensitive attributes*.

Fairness Mitigation

It is important not only to measure the fairness, but also to address their unfairness. Mitigation techniques aim to reduce or eliminate bias and unfairness of the models [54] [55]. They are usually divided into three categories:

- **Pre-processing**. It consists of modifying the training data before the model is trained. It involves methods like re-weighting (i.e. changing the weights of the samples), or modifying features that are correlated with protected attributes.
- **In-processing**. They include fairness constraints in the training process. Some widely-used in-processing methods are adversarial debiasing, and exponentiated gradient.
- **Post-processing**. It consists of modifying the output of the model to ensure fairness. It encompass techniques to adjust the decision thresholds to ensure that the TPR, FPR, or both are equal across the groups of the protected attributes.

1.2.5 Feature Importance

Up to this point, attention has primarily been directed towards the predictive performance and fairness of the models. However, to understand how the models takes decisions we must go further and analyse the explainability of the models, which involves inspecting the most important features that guide their predictions [56] [57].

SHAP, which stands for SHapley Additive exPlanations, is a method used to explain individual model predictions based on the Shapley values [58]. Shapley values are computed by measuring the difference in the model output between including a feature and not including it. The Shapley value for a given feature is calculated through the average of its contributions to all possible feature combinations (i.e., all possible subsets of other features). Also, the average contribution is weighted by the number of features that are not present in the subset.

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \Big(f_{S \cup \{i\}}(x) - f_S(x) \Big) \tag{1.19}$$

where:

- N is the set of all features.
- S is any subset of features excluding i.
- $f_S(x)$ is the model prediction using the features in the subset S.
- $f_{S \cup \{i\}}(x)$ is the model prediction using the features in the subset S, and feature i.
- $|S|$ is the number of features in subset S.
- $|N|$ is the total number of features.
- $|S|!$ is the number of permutations of features in subset S.
- $|N| - |S| - 1)!$ is the number of permutations of features not present in subset S, also excluding feature i.

An interesting characteristic of the Shapley values is that they provide both the magnitude and direction of the feature contribution, which can lead to gain insights into how higher or lower feature values affect model predictions.

1.3 Structure of the document

This thesis is organized into seven main chapters. The description of each chapter is as follows:

- **Motivation and Background**. In this chapter, the whole introduction of the thesis was explained, including its motivation and theoretical framework.
- **Objectives**. In this chapter contains the main objectives and the subobjectives of the thesis.
- **State of the art**. This chapter contains the literate review of the most recent and cutting edge methods for diagnosis prediction with machine learning, as well as cardiovascular disease diagnosis and prognosis of machine learning models using radiomics and fractal-based derived from cardiovascular medical imaging.
- **Design and development**. In this chapter, the methodology is explained in depth, including the subcohort creation, the used feature sets, the replication of state of the art radiomics results, an overview of the machine learning pipeline, the pre-processing steps, the used machine learning models, the hyperparameter tuning, and the model evaluation.
- **Experiments and results**. In this chapter, the main results of the experiments are shown.
- **Discussion and conclusion**. This chapter addresses the discussion of the results, the assumptions and limitations of the work, and the final conclusions.
- **Future work**. This chapter focuses on the future work to be done after the contributions of this thesis.

Chapter 2

Objectives

The primary aim of this thesis is to assess the efficacy and fairness of fractal-based features derived from cardiac magnetic resonance imaging (CMR) in predicting the diagnosis and prognosis of complex cardiovascular diseases, utilizing advanced machine learning techniques.

Specifically, the target cardiovascular diseases are: angina, arrhythmia, heart failure, myocardial infarction, stroke, and valvular heart disease.

To achieve the primary aim of this thesis in a comprehensive way, the following subobjectives must be accomplished:

- Develop an automated machine learning pipeline for the diagnosis and prognosis of cardiovascular diseases using image-based features extracted from cardiac magnetic resonance images.
- Evaluate the performance of the developed pipeline by replicating the results of state-of-the-art radiomics-based machine learning studies to ensure the consistency of data pre-processing and machine learning pipeline.
- Evaluate and compare the predictive performance of the machine learning models using fractal-based features against those employing radiomics-based features.
- Examine the potential added value of integrating fractal-based features with traditional CMR indices or radiomics features in enhancing diagnostic and prognostic models.
- Assess and compare the fairness of the best-performing models based on fractal-based and radiomics-based features in terms of sensitive attributes such as sex, age, overweight and high cholesterol.

- If necessary, mitigate the potential biases committed by the best-performing models.
- Study the feature importance of the best-performing fractal models to guide future research into such underexplored features in cardiovascular research.

Chapter 3

State of the art

3.1 Machine learning with CMR imaging data

Machine learning techniques have been extensively employed to detect and diagnose several pathologies [59] [60] [61]. Some of these studies specifically focus on cardiovascular diseases and use CMR imaging data. Among them, the majority concentrate on segmentation [62], which is crucial for computing conventional CMR indices. For instance, Davies et al. (2022) [63] used Convolutional Neural Networks (CNNs) based on the U-Net architeture to segment the left ventricular blood volume and myocardium in CMR images. Their model demonstrated superior performance in terms of scan-rescan precision (i.e. ability to reproduce similar results when analyzing images from two separate scans of the same patient) compared to the segmentation provided by three medical experts.

Although to a lesser extent, some papers have also been published on the use of machine learning techniques to predict cardiovascular disease diagnoses using CMR imaging data [62]. Swift et al. (2020) [64] used tensors with both spatial and temporal information of the CMR data and applied multilinear principal component analysis to reduce its dimensionality. Then, they performed feature selection using Fisher's discriminant ratio, and used support vector machine (SVM) and Logistic Regression (LR) classifiers to diagnose pulmonary arterial hypertension. The proposed pipeline obtained 0.92 in the area under the receiver operating characteristic curve (AUC-ROC), with the SVM model being slightly better than conventional CMR indices.

3.2 Machine learning with radiomics-based features from CMR images

Although still an emerging field, some researchers have proposed models that incorporate radiomics-based features from CMR images. Durmaz et al. (2022) [65] predicted major adverse cardiac events using radiomics features derived from late gadolinium enhancement (LGE) and cine images. Several radiomics features were extracted, including first-order, shape, and texture features, and they were selected using inter-class correlation (ICC) coefficients and LASSO. Conventional CMR indices as well as other clinical information (e.g., age, sex, BMI, etc.) were also considered, creating various combinations of features sets. They employed repeated random sampling and used several ML models: adaptive boosting (AdaBoost), k-nearest neighbors (k-NN), naive Bayes (NB), neural network (NN), stochastic gradient descent, and SVM. The highest performance was obtained by the neural network with all the feature sets (radiomics, conventional CMR indices, and clinical features), showing an AUC-ROC of 0.965.

Avard et al. (2022) [66] extracted and used first-order, shape, and texture radiomics features to detect myocardial infarction from non-contrast cine CMR images. They performed feature selection by ranking the features with multiple support vector machine recursive feature elimination (MSVM-RFE), computing the Spearman correlation between features, and removing the ones which were highly correlated. As the machine learning models, they used logistic regression (LR), linear discriminant analysis, quadratic discriminant analysis, extra tree, random forest (RF), AdaBoost, k-NN, NB, SVM and NN. The best-performing model was the LR, with an AUC of 0.93 and a F1 Score of 0.90.

Cetin et al. (2020) [24] considered various cardiovascular factor groups from a UK Biobank sample, including hypertension, diabetes, high cholesterol, current smoker, and previous smoker. They extracted shape, signal first-order and texture-based radiomics features from manually annotated segmented CMR cine images. The features were selected using sequential forward feature selection (SFFS), being the machine learning models SVM, RF, and LR. LR with L1 regularization achieved the highest performance in terms of AUC-ROC, yielding 0.80 for diabetes, 0.72 for hypertension, 0.71 for high cholesterol, 0.68 for current smoker, and 0.63 for previous smoker. The performance metrics surpassed conventional CMR indices in all the groups. In addition, radiomics models showed superior performance, and the difference was statistically significant in every cardiovascular risk factor group, except for current smokers, as assessed by the McNemar's test.

Pujadas et al. (2022) [25] predicted incident atrial fibrillation (AF), heart failure (HF), myocardial infaction (MI), and stroke, in a sample from the UK Biobank. They tried different combinations of feature sets, which included vascular risk factors, conventional CMR indices, and radiomics features, that consisted of first-order, shape, and texture features. The features were selected using SFFS, and the chosen classifier was SVM. The evaluation involved a two-level cross-validation process, with the inner cross-validation dedicated to tuning the hyperparameters, and the outer one for the final assessment. Radiomics features were incorporated into the feature set of the best-performing models for all diseases tested, suggesting the potential benefit of including these features alongside VRF, CMR indices, or both.

Study	**Prediction**	**Feature selection**	**Model**	**Main finding**
Durmaz et al. (2022) [65]	Major adverse cardiac events	Inter-class correlation, LASSO	NN	High performance (AUC-ROC: 0.965)
Avard et al. (2022) [66]	Myocardial infarction	MSVM-RFE, Spearman correlation	LR	High performance (AUC-ROC: 0.93, F1: 0.90)
Cetin et al. (2020) [24]	Hypertension, diabetes, high cholesterol, current and previous smoker	SFFS	LR (L1)	Superior to conventional CMR indices in all diseases
Pujadas et al. (2022) [25]	Incident AF, HF, MI and stroke	SFFS	SVM	Potential of including radiomics with VRF and conventional CMR indices

Table 3.1: Summary of studies utilizing machine learning with radiomics-based features extracted from CMR images.

3.3 Fractal-based features and cardiovascular diseases

Few studies have employed fractal analysis in CMR. Wang et al. (2020) [67] examined the prognostic value of LV trabecular complexity measured with fractal analysis of CMR cine images in patients with hypertrophic cardiomyopathy (HCM). They set the primary endpoint to all-cause mortality and aborted sudden cardiac death (SCD), and the secondary endpoint to readmission due to HF. They extracted the endocardial boundaries to compute the fractal dimensions (FD) with the box-counting approach. Having a higher maximal apical FD was associated with both the primary and secondary endpoints. Considering typical cardiovascular predictors and the percentage of LGE in a multivariate Cox model, LGE maximal apical FD

was an independent predictor for both primary and secondary end points. Internal validation, done through bootstrapping, showed that left ventricular (LV) maximal apical FD predicted the primary endpoints with a mean AUC-ROC of 0.70. More recently, Jiang et al. (2023) [68] have shown that considering also the trabecular complexity of the right ventricle along with the LV in the could improve the model, achieving a C-index of 0.864 for the primary endpoint (SCD) and 0.877 for the secondary endpoint (readmission due to HF).

Similarly, even more limited research has been conducted on the application of machine learning using fractal-based features in cardiovascular imaging. Firouznia et al. (2021) [69] predicted AF recurrence following catheter ablation using fractal features extracted from segmented CT scans. This included the extraction of shape-based features by calculating 1D, 2D, and 3D fractal dimensions with the box-counting method, as well as texture features representing the texture variations of the CT intensities. Different feature sets were considered, including the fractal and morphological properties of the left atrium (LA), pulmonary veins (PV), LA myocardial tissue (LAM), and all the mentioned feature types together. Additionally, clinical features such as BMI, hypertension, LA volume, ablation type, etc., were also considered. Fractal features of LA, LAM, and the feature set with all the fractal features were associated with the risk of AF recurrence. In the machine larning experiment, features were selected by using the 5 features with the lowest p-value in the Wilcoxon rank-sum test, and the chosen classifier model was a RF. The highest performance was obtained by the model that fractal feature types, achieving an AUC-ROC of 0.81, which was superior to that of the clinical features model. Combining both models (considering all defined feature types and clinical features) showed the best overall predictive performance, with an AUC-ROC of 0.87.

The only study available about CMR is by Kurzendorfer et al. (2018) [70], which employed fractal analysis to segment myocardial scars in LGE magnetic resonance images. Through the box-counting approach, the fractal dimensions of the boundaries were extracted from the images, which were previously binarized by a two-threshold binary decomposition. Additionally, the size of the binary object and the mean gray value were included as features. The chosen classifier, a RF, predicted a probability for each pixel. The scar quantification approach consisted of converting this probability to binary through a threshold, using connected component analysis to discard small and weakly-connected components, and applying morphological closing to seal small holes. Their approach yielded a mean Dice of 0.64, surpassing the x-fold standard deviation method.

Study	Task	Fractal Computation	Performance	Main Finding
Wang et al. (2020) [67]	Prognosis in HCM with CMR	Box-counting	AUC-ROC 0.70 with Cox regression (max. apical FD)	LV maximal apical FD was an independent predictor for all-cause mortality and SCD.
Jiang et al. (2023) [68]	Prognosis in HCM with CMR	Box-counting	C-index 0.864 with Cox regression (FD + VRF + LGE)	Improved model by including RV trabecular complexity, showed significant prediction for SCD.
Firouznia et al. (2021) [69]	Prognosis in AF with CT	Box-counting	AUC-ROC 0.87 with RF (fractal + clinical features)	Fractal features were associated with the risk of AF recurrence, and showed a higher performance than clinical features.
Kurzendorfer et al. (2018) [70]	Segmentation of myocardial scars with CMR	Box-counting	Dice 0.64 with RF	Effective segmentation of myocardial scars in LGE MR images, surpassing standard methods.

Table 3.2: Summary of studies employing fractal analysis in cardiovascular imaging.

Chapter 4

Design and development

4.1 Dataset

The **UK-Biobank** [71] is a biomedical database from over half a million participants residing in the United Kingdom. It includes a diverse range of data types, such as genetic profiles, medical imaging results, biochemical analyses, and physical activity metrics. The sample used in this thesis comprises 32,003 subjects from the UK-Biobank for whom CMR imaging data was available.

In this thesis, both diagnosis and prognosis of the cardiovascular diseases were considered. In the **diagnosis** scenario, the criteria for the positive case was that the participant was diagnosed with a given cardiovascular disease up to the imaging date (i.e. the date when the cardiac magnetic resonance was performed). For **prognosis**, participants were considered a positive case if they were diagnosed with the target cardiovascular disease after the imaging date up to 10 years of follow-up. However, not all participants joined into the UK Biobank cohort at the same time, so the follow-up for some patient that joined later was shorter. In those cases, November 2023 -the latest date for which the data was available- was used as censoring date. Figures 4.1 and 4.2 illustrate some cases of patients that represent a positive or a negative case for the diagnosis and prognosis scenario, respectively.

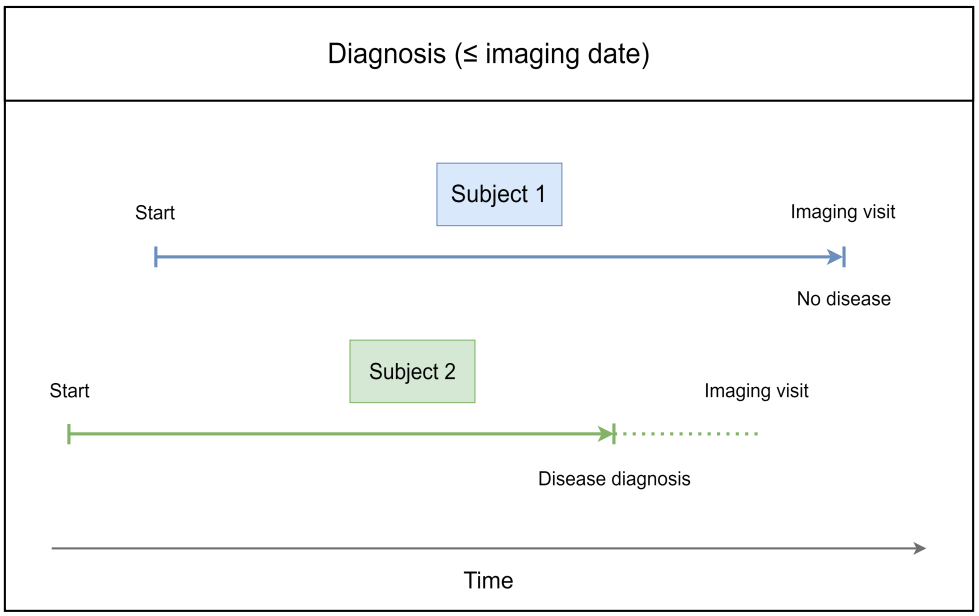


Figure 4.1: Diagnosis scenario. In the illustrated example, subject 2 was diagnosed with the target disease before the imaging date, so is considered a positive case, whereas subject 1 was not, thus being a control.

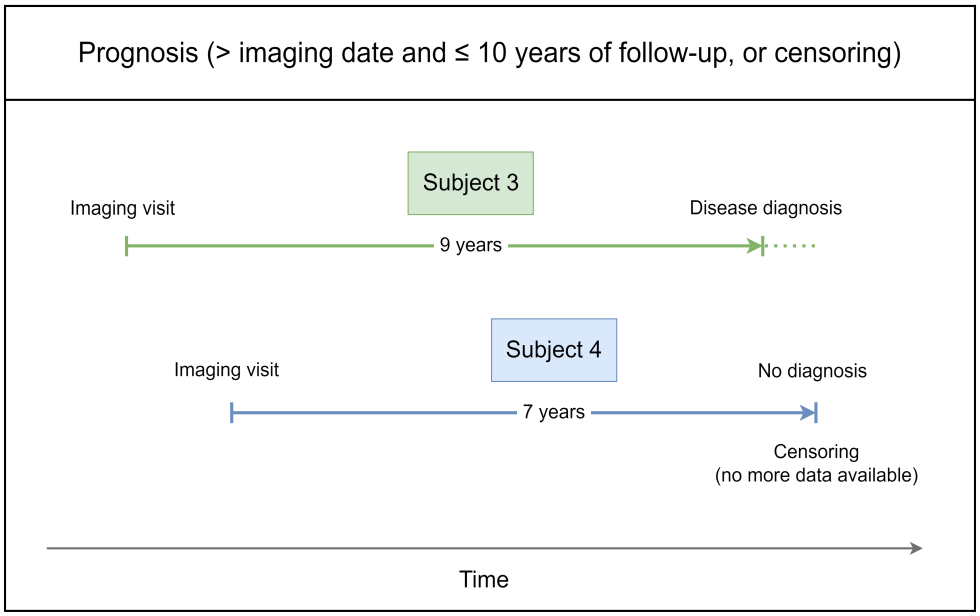


Figure 4.2: Prognosis scenario. In the illustrated example, subject 3 was diagnosed with the target disease during the follow-up period, so is considered a positive case. Subject 4 was not diagnosed with the disease during the 6 years that he or she could be followed. Since the maximum follow-up could not occur due to a lack of more recent information, subject 4 would be considered a control case.

Diseases were identified according to the International Classification of Diseases, tenth revision (ICD-10) codes. Medical experts defined component outcomes grouping some cardiovascular individual diseases. In particular, six component outcomes were considered: angina, arrhythmia, heart failure, myocardial infarction, stroke, valvular heart disease. The correspondence between ICD-10 codes and the component outcomes can be checked in Annex A. The control group consisted of participants who were not diagnosed with the target disease during the specified time period. However, these individuals could suffer from other diseases and could not be completely free from other cardiovascular diseases.

The subcohort with the positive and negative cases for each disease was created from the files of the whole cohort provided by the UK-Biobank. Specifically, three data fields where used:

- **Data field 53**, which contains information about the **date of attending** to the assessment centre. Among its instances, the one of interest was the date of the imaging visit, since it represents the upper and lower time limits for diagnosis and prognosis settings.
- **Data field 41270**, which contains the **ICD-10 codes** and the description for each disease diagnosed to the participants.
- **Data field 41280**, which contains the **date of the first diagnosis** for disease of the participants. This was used to determine whether a participant was considered a positive case, if it also occurred within the time window for diagnosis or prognosis.

4.2 Image-based features

In this section, the utilized feature set will be described, including conventional CMR indices, radiomics-based features and fractal-based features.

4.2.1 Conventional CMR indices

The following conventional CMR indices were employed in this thesis: end-diastolic volume, end-systolic volume, stroke volume, ejection fraction and LV mass. All measurements but the ventricular mass (only used for the left ventricle), were utilized for both the left and right ventricles. Moreover, CMR indices were normalized by the body surface area (BSA) to account for individual body size differences.

4.2.2 Radiomics Features

For a given anatomical part (right ventricle, left ventricle, or myocardium) and phase (end-diastole or end-systole), 105 features were extracted, including 13 shape features, 18 first-order features, and 74 second-order features (23 from GLCM, 16 from GLSZM, 16 from GLRLM, 5 from NGTDM, and 14 from GLDM). Considering all six combinations of anatomical parts and phases (RV-ED, RV-ES, LV-ED, LV-ES, MYO-ED, MYO-ES), this results in a total of 630 features.

4.2.3 Fractal Features

In this thesis, a total of 12 fractal-based features were used, which consists of the fractal dimension and the lacunarity for each of the phases (end-systole and end-diastole) and anatomical parts (left ventricle, right ventricle, or myocardium). It is important to note that for this master's thesis, the fractal related computations were not performed, but the extracted fractal-based features were provided instead.

Fractal dimension through differential box-counting

The fractal dimensions were computed using the **differential box-counting** method [72], which considers both spatial layout and intensity variations. For each grid size, the volume was uniformly divided into non-overlapping blocks of identical dimensions. To accurately represent the intensity variation of each block, the number of required boxes was computed considering the maximum and minimum intensity values as well as a scaling factor to adapt the granularity of grey levels in the different scales.

$$N_\delta(i,j,k) = \left\lceil \frac{\max(I) - \min(I)}{\delta'} + 1 \right\rceil, \tag{4.1}$$

where:

- $\max(I)$ and $\min(I)$ are the maximum and minimum grey level intensities in the block (i,j,k).
- δ' is the adjusted grid size, calculated based on δ and the total number of grey levels, computed as $\delta' = \frac{\delta \cdot \text{Total Grey Levels}}{\text{Dimension of the Image}}$.

- δ is the grid size before being adjusted by the gray levels, which is a divisor of the dimensions of the image.

A series of offsets were utilized to optimize the count of observable grey level points (i.e., the range of intensity values from the minimum to the maximum within each grid). This was achieved by shifting the blocks of the grid along every axis in the positive direction, or in the negative direction if the shift exceeded the original volume boundaries. Offsets were chosen by maximizing the number of discernible grey levels, enhancing the accuracy of the box count. The maximum count of observable data points was then plotted on a log-log scale against the inverse of the grid sizes. Finally, a least square linear fit was applied to these points to compute the fractal dimension. Thus, the fractal dimension is estimated with the slope of the linear fit of $\log(N_\delta)$ against $\log(\delta^{-1})$. Figure 4.3 illustrates the fractal estimation of a Sierpinski triangle using the box-counting method.

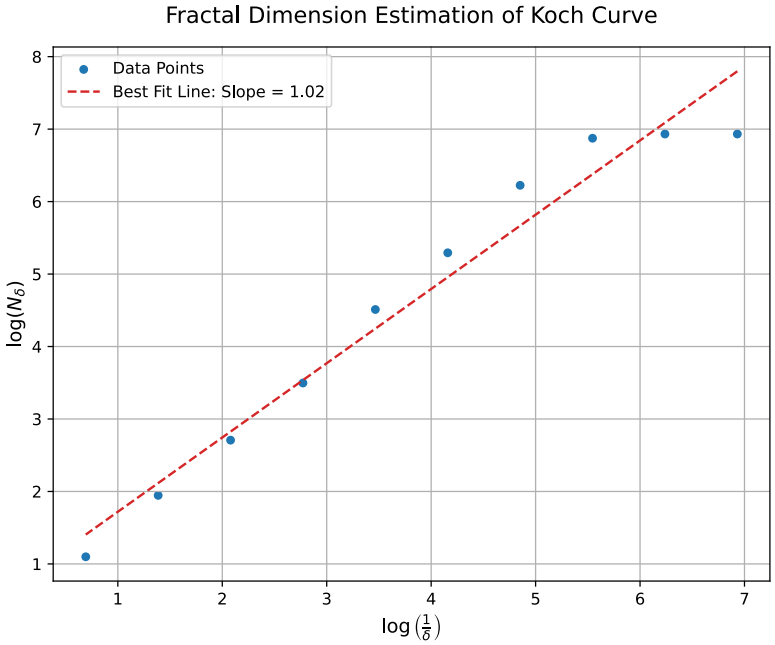


Figure 4.3: Estimation of the fractal dimension for the Koch curve with 5 iterations using the box-counting method. The plot illustrates the relationship between the logarithmic of the box count (N_δ) and the logarithm of the inverse of the box size (δ). The least squared linear fit (indicated with a dashed red line) represents the estimation of the fractal dimension.

Texture assessment through lacunarity

The computation of lacunarity used an adaptation of the **Gliding-Box Lacunarity algorithm** for three-dimensional structures [73]. Unlike the box-counting method, which used non-overlapping blocks, the computation of lacunarity was conducted by dividing the image into overlapping blocks. For each block, the mass $m_\delta(i,j,k)$ was calculated based on the minimum and maximum grey levels in that block.

$$m_\delta(i,j,k) = \begin{cases} \left\lceil \frac{\max(I_{i,j,k}) - \min(I_{i,j,k}) + 1}{\delta'} \right\rceil, & \text{if } \max(I_{i,j,k}) \neq \min(I_{i,j,k}) \\ 1, & \text{if } \max(I_{i,j,k}) = \min(I_{i,j,k}) \end{cases}, \tag{4.2}$$

where:

- $\max(I_{i,j,k})$ and $\min(I_{i,j,k})$ are the maximum and minimum grey level intensities in the block i,j,k.
- δ' is a scaling factor for the grid size, which adapts to the granularity of the grey levels.

Then, the first and second-order moments of the mass distribution of all blocks were computed, which represent the average mass per block and the average squared mass per block, respectively.

$$Z_Q^{(1)} = \frac{\sum_{i,j,k} m_\delta(i,j,k)}{n}, \quad Z_Q^{(2)} = \frac{\sum_{i,j,k} m_\delta(i,j,k)^2}{n} \tag{4.3}$$

where:

- $m_\delta(i,j,k)$ is the mass of each block.
- n is the total number of blocks in the grid.

Lacunarity was then calculated using these moments.

$$\Lambda(\delta) = \frac{n \cdot Z(2)_Q}{(Z(1)_Q \cdot n)^2} \tag{4.4}$$

where:

- $Z(1)_Q$ and $Z(2)_Q$ are the first and second moments of the mass distribution.

Finally, the average lacunarity across all scales was computed to provide a single value that reflects the whole lacunarity of the structure.

4.2.4 Feature Set Construction

An ablation experiment was conducted to evaluate the predictive performance of all possible combinations of feature sets, which are conventional CMR indices, fractal, and radiomics-based features. This approach allows to check if there is a benefit from including fractal or radiomics-based features along with traditional CMR features, or by combining fractal and radiomics-based features. In all feature sets, age and sex were included, as they provide essential information about the context of the subject. For example, a specific radiomics or fractal feature value could have a very different meaning for a younger male compared to an older female.

4.3 Replication of state-of-the-art radiomics results

To assess the performance of the developed machine learning pipeline, the first step was compare against the state-of-art in radiomics-based ML. More precisely, the results of the Pujadas et al. (2022) [25] study were replicated, which utilized radiomics-based features to prognosticate various cardiovascular diseases using the same study sample from the UK Biobank. Specifically, the entire Supplementary Table 6 was reproduced, which included the performance of prognosis predictions for atrial fibrillation (AF), myocardial infarction (MI), and heart failure (HF). The control group consisted of participants who were not diagnosed with any cardiovascular diseases during the follow-up period.

Conventional CMR, vascular risk factors, and radiomics features, as well as any combination of them, were employed as feature sets. A total of 262 radiomics features were utilized, including shape, first-order, and texture features. The same machine learning pipeline as the original study was employed, except for the feature selection method. Instead of using sequential forward feature selection (SFFS), which required an excessively long execution time (in the order of days), features were selected based on model weights (which reflect their importance), choosing those with absolute importance equal to certain threshold.

The machine learning model utilized was a Support Vector Machine (SVM). As figure 4.4 illustrates, model assessment was conducted using two-level cross-validation (CV): the inner CV for hyperparameter tuning and the outer CV for testing the model. Cross-validation was performed using stratified 5-fold CV, so that the same proportion of classes of the target variable as the whole training set was kept in each split.

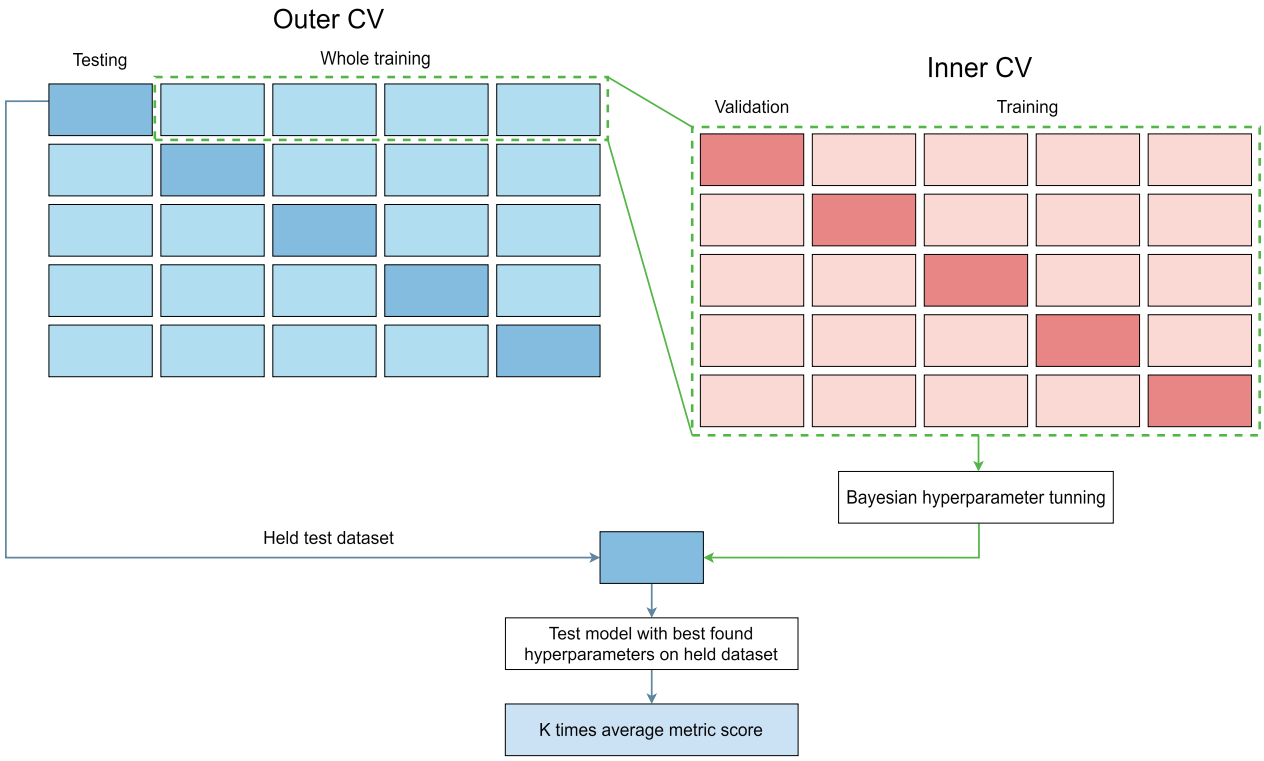


Figure 4.4: Two-level cross-validation scheme. Adapted from Rauseo et al. (2021) [74]. The outer cross-validation is used for testing, whereas the inner cross-validation is employed for hyperparameter tuning.

Evaluation metrics included accuracy, sensitivity, specificity, and AUC-ROC, with the latter serving as the primary performance metric. Parameters were optimized through Bayesian optimization, focusing on maximizing the AUC-ROC. The optimization covered the regularization parameter (C), the kernel coefficient, and the choice of kernel function. Specifically, the following hyperparameter ranges were used:

- C: [0.01, 1000]
- Gamma: [0.0001, 2.0]
- Kernel: Linear, radial basis function (RBF), polynomial, and sigmoid.

Additionally, the hyperparameter C and the importance threshold used for feature selection were also included in the optimization.

- C (feature selection): [0.01, 1000]
- Threshold (proportion of the mean importance): 0.1, 1

4.4 Overview of the Machine Learning Pipeline

The general machine learning pipeline consists of three parts: hyperparameter tuning, fitting the model with the optimized hyperparameters, and the final testing.

Similar to the evaluation method illustrated in figure 4.4, the pipeline involved a two-level cross-validation, with the outer CV for testing and the inner CV for hyperparameter tuning. First, we should define what training, validation, and testing data mean in this context, which is not as straightforward as a regular train-test-validation split.

- **Outer training data.** It represents all data but the held fold used for testing. In the traditional train-test-validation split method, it would be analogous to the training plus validation sets.
- **Inner training data.** It represents the data in the inner cross-validation not in the held set used for validation during hyperparameter tuning. In the train-test-validation analogy, it would be the training set.
- **Testing data.** It represents the data of the held fold of the outer CV used for testing. It would be similar to the testing set of the train-test-validation method.

In the hyperparameter tuning phase, the inner training data was preprocessed by fitting and transforming the data with a scaler, an undersampler, and a feature selector. Then, the hyperparameters were chosen and the classifier was trained using these hyperparameters. The choice of the hyperparameter values was based on Bayesian optimization through the Optuna framework. Following this, the same preprocessing steps that were applied to the inner training data were also applied to validation (i.e., inner held fold) data. However, under-sampling was not applied to the validation data to avoid modifying the real-world distribution of the classes. Subsequently, the model performance was assessed using 3-fold cross-validation. The described process was performed for each of the Optuna trials, and

the hyperparameter value combination that maximized the mean performance metric in the cross-validation was extracted.

Once the hyperparameter tuning procedure was finished, all pre-processing steps were fitted and applied to the outer training set, which included both training and validation data. The classifier was then fitted with the optimized hyperparameters on the pre-processed outer training (i.e., training plus validation) data. In this way, the model could be trained with all the available non-testing data (i.e., all data but the held fold), while data leakage was not committed. This is usually referred to as *refitting* the model.

Finally, the last step corresponds to testing. The same pre-processing transformations that were done to the outer training data were applied to the testing (i.e., held fold) data, except for the sampling.

An overview of the proposed approach is provided in figure 4.5.

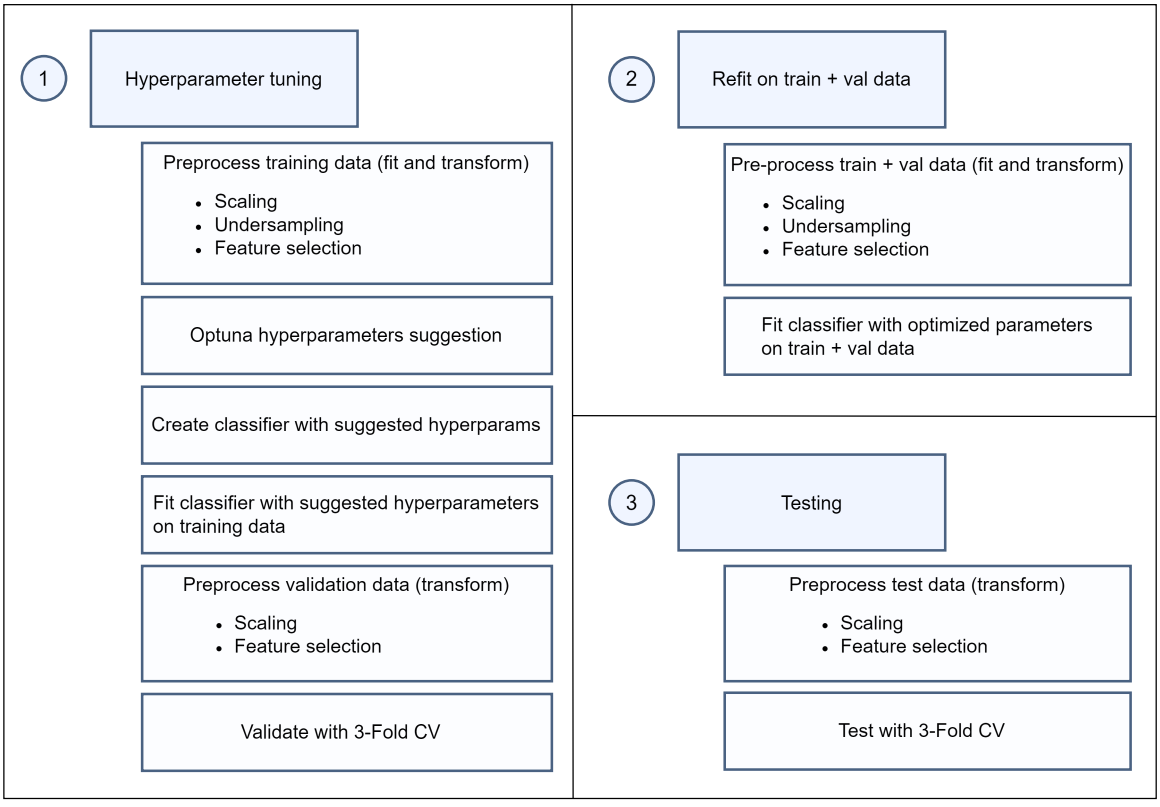


Figure 4.5: Steps of the general machine learning pipeline followed in this thesis.

4.5 Pre-processing

4.5.1 Encoding

All the nominal variables used were binary and were label-encoded, which occupies less memory compared to one-hot encoding by reducing the size of the feature matrix [75].

4.5.2 Scaling

Several features types were used, and most of them were in very different scales. In this context, scaling features is important because otherwise, some models that are sensitive to the scale of the features (like SVM), could give more importance to features that have wider scales. Additionally, scaling features could lead to faster convergence in models that use optimization based on the gradient descent (like XGBoost), since it could be quicker for the algorithm to find the global minimum [76]. The method used for scaling the features was Min-Max, which scales the feature values between 0 and 1 by using their minimum and maximum.

$$x_{\text{scaled}} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{4.5}$$

4.5.3 Undersampling

The participants of the UK Biobank who underwent cardiac magnetic resonance represent the whole population and were not subjected to imaging because they had signs of cardiovascular disease, but because they were asked for research purposes. As they represent the general population (or even healthier for the *healthy volunteer* effect), the difference in the number of subjects in the diseased and non-diseased groups is very large, with the control group much larger, which results in a significant class imbalance.

To prevent models from learning more patterns for the most frequent classes, the classes were forced to be balanced in the training set. This was achieved through random undersampling, which randomly selects some samples from the most frequent category to match the number of the less frequent one. The choice of undersampling instead of oversampling is due to the huge difference between the classes [77], with the minority class being less than 3% of the majority one in all cases, which would have resulted in most of the samples being synthetic for the minority class. It is worth to mention that undersampling was not utlized for the balanced

random forest, since it internally performs sampling to balance the classes, as explained in section 1.2.2.

4.5.4 Feature Selection

It was decided to apply feature selection due to the large number of features in the radiomics feature set (630 features). To make the comparison between fractals and radiomics as fair as possible, feature selection was applied to all feature sets regardless of their size, employing the exact same machine learning pipeline (including feature selection).

The used feature selection method consisted of selecting features based on their model importance weight or coefficient [78]. More specifically, features were selected if their absolute weight was greater or equal to a given threshold.

$$\text{Selected Features} = \{x_i \mid |w_i| \geq \text{threshold},\ i = 1, 2, ..., n\}, \tag{4.6}$$

where:

- x_i is the ith feature in the dataset.
- w_i is the model importance weight or coefficient for the ith feature.

The threshold was computed as a proportion of the mean importance of all features.

$$\text{threshold} = \alpha \times \frac{1}{n} \sum_{i=1}^{n} |w_i|, \tag{4.7}$$

where:

- α is a predefined value to define the threshold.

4.6 Machine Learning Models

The used machine learning models were SVM, XGBoost, LightGBM, AdaBoost, random forest and balanced random forest. In the case of balanced random forest, its sampling strategy involved a combination of both oversampling the minority classes and undersampling the majority one.

4.7 Hyperparameter Tuning

4.7.1 Optimization Method

The optimization of the hyperparameters was addressed by using the Tree-structured Parzen Estimator (TPE) implementation of the Optuna framework [51]. For each of the 200 trials, the model performance with the hyperparameter values suggested by TPE was evaluated in terms of its mean balanced accuracy obtained in the inner 3-Fold cross-validation. The selected hyperparameter values were those that maximized the mean balanced CV accuracy.

4.7.2 Hyperparameter Ranges

For feature selection, the number used to compute the importance threshold was optimized within a range from 0.1 to 1 as a proportion of the mean feature weight of the model. This allowed to select both a small and a large number of features, taking into account the different sizes of the feature set.

Regarding SVM, a wide range for the C hyperparameter was used to control the trade-off between the loss and the margin. Also, several kernels were tried, with a special focus on non-linear kernels, which can capture more complex patterns. The *gamma* hyperparameter in the case of radial basis function (RBF) kernel was optmized as well.

With respect to the gradient boosting models, similar hyperparameter ranges were used whenever possible. The tendency to overfitting that decision trees usually have if they have no constraint [79] when they are built was addressed through several hyperparameters. For example, the maximum depth of the decision trees was controlled with a maximum value. Similarly, a minimum limit in the number of leaf nodes samples and a minimum gain in the splits were employed. The learning rate was the same for XGBoost and LightGBM, but it

was higher for AdaBoost, as in that model it does not represent a regular learning rate, but the weight of the base estimators in the final model, for which not such low values are usually used [80].

In addition to the hyperparameters shared between the gradient boosting and the random forest models, in the latter the number of features was also limited using the squared root or logarithm of the total number of features.

Moreover, model complexity was restricted in all ensemble models with a maximum number of estimators, using an upper limit of 100.

The exhaustive list of hyperparameter ranges considered for each model can be checked in Annex B.

4.8 Model Evaluation

Model evaluation was conducted in a comprehensive manner from different perspectives, considering both the predictive performance and the fairness of the models. In addition, to ensure model explainability, feature importance was analysed.

4.8.1 Performance Evaluation

Evaluation method

For evaluating the models, a two level cross-validation was used, employing the inner CV for hyperparameter tuning and the outer CV for testing the model. Due to the long execution time and the high number of settings tested (all combinations of feature sets and several diseases for diagnosis and prognosis), 3 folds were chosen for both CVs.

Performance Metrics

Usually, in real-world scenarios, like cardiovascular disease diagnosis and prognosis, the datasets are imbalanced. In this context, providing the regular accuracy of the model could be misleading, since the model could achieve a high accuracy just by predicting always the majority class. Instead, metrics that account for the number of observations or average the results (like balanced accuracy) provide a more complete view of the real predictive performance of the model [81].

The confusion matrix is a table that provides information about the number or rate of the predictions where the model and the real labels agree. More interestingly, it also provides information about which categories have been incorrectly classified in the wrong predictions.

In binary classification, there are two cases of agreement and disagreement between the model predictions and the actual labels:

- **True Positives** (TP). The number of cases in which the model correctly classifies the positive class.
- **True Negatives** (TN). The number of cases in which the model correctly classifies the negative class.
- **False Positives** (FP). The number of cases in which the model prediction is the positive class but the actual label is the negative class.
- **False Negatives** (FN). The number of cases in which the model prediction is the negative class but the actual label is the positive class.

Table 4.1 shows the scheme of the confusion matrix for binary classification.

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positives (TP)	False Negatives (FN)
Actual Negative	False Positives (FP)	True Negatives (TN)

Table 4.1: Confusion matrix for binary classification

The main performance metric used in this thesis was the balanced accuracy. **Balanced accuracy** is computed as the mean of the sensitivity or recall (true positive rate) of the classes. By averaging the true positive rate of both classes using equal weights (regardless of the sample size), it provides a more equitable measure to evaluate model performance compared to regular accuracy. For binary classification, it is computed as:

$$\begin{aligned} \text{Balanced Accuracy} &= \frac{1}{2}\left(\text{Recall}_{\text{class 1}} + \text{Recall}_{\text{class 2}}\right) \\ &= \frac{1}{2}\left(\frac{TP_{\text{class 1}}}{TP_{\text{class 1}} + FN_{\text{class 1}}} + \frac{TP_{\text{class 2}}}{TP_{\text{class 2}} + FN_{\text{class 2}}}\right) \end{aligned} \tag{4.8}$$

4.8.2 Statistical Testing

Comparative Performance Analysis with Wilcoxon Signed-Rank Test

The main objective of the current study is to compare the performance of the radiomics and fractal models. Since the method used to test the models is an outer cross-validation, this was addressed by performing statistical testing, comparing the performance of the best radiomics and fractal models between the CV folds for each disease. More concretely, the Wilcoxon Signed-Rank test was applied, which is a non-parametric test that compares the ranks of differences between paired samples [82]. For each fold, the **signed ranks** were computed, which consider the magnitude of the differences in the balanced accuracy scores, along with their sign.

$$W_i = \text{sign}(X_i - Y_i) \times \text{rank}(|X_i - Y_i|), \tag{4.9}$$

where:

- X_i and Y_i are the balanced accuracy scores from the two models in the i-th fold.
- $\text{sign}(X_i - Y_i)$ is the sign of the difference between the balanced accuracy scores of the models in the i-th fold.
- $\text{rank}(|X_i - Y_i|)$ ranks the balanced accuracy scores based on the absolute value of the differences.

Then, the **W statistic** was computed, which is the minimum between the sum of ranks with positive signs and those with negative signs.

$$W = \min(W^+, W^-), \tag{4.10}$$

where:

- $W^+ = \sum_{W_i>0} W_i$ is the sum of the positive signed ranks.
- $W^- = \sum_{W_i<0} W_i$ is the sum of the negative signed ranks.
- $W_i = \text{sign}(D_i) \times \text{rank}(|D_i|)$ is the signed rank for each observation.
- D_i is the difference between the balanced accuracy scores of the models in the i-th fold.

The null hypothesis of the Wilcoxon Signed-Rank test holds that the median difference of the paired scores is equal to zero, and thus there are no differences in the performance of the models. In contrast, the alternative hypothesis holds that the median difference is not zero, suggesting that the performance of both models is dissimilar.

- $H_0 : \text{median}(D_i) = 0$ where $D_i = X_i - Y_i$ represents the difference in performance scores between the two models for each fold.
- $H_1 : \text{median}(D_i) \neq 0$.

Once the W statistic was computed, its associated p-value was calculated as well. If the p-value is greater than the significance value ($\alpha = 0.05$), there is not enough evidence to reject the null hypothesis, and it is considered that there are no significant differences between the models. Otherwise, the null hypothesis is rejected in favor of the alternative, assuming the differences in performance are significant.

4.8.3 Fairness

Fairness Metrics

The used metrics to evaluate the model fairness were demographic parity, equal opportunity difference, and average odds difference, which are some of the most commonly used fairness metrics [83] [84].

Demographic Parity

Demographic Parity (DP) is a condition that is met when the probability of the favorable outcome is the same between all groups defined by the protected attributes. In a fair model, the probability of positive prediction should be equal regardless of the group to which a participant belongs. Demographic parity can be expressed with an equality based on the conditional probability of having a positive outcome if we know a participant belongs to a certain group:

$$P(\hat{Y} = 1|A = a) = P(\hat{Y} = 1|A = b), \tag{4.11}$$

where:

- $\hat{Y}$ is the predicted outcome.
- A is the protected attribute (e.g., sex).
- a and b are different groups of the protected attribute (e.g., male and female).

As a quantitative measurement to assess fairness, the difference in the probabilities is used, with 0 being the ideal case and 1 the most unfair case possible.

Equal Opportunity Difference

The Equal Opportunity Difference (EOD) is similar to the demographic parity but it focuses on the sensitivity (true positive rate). It is measured as the difference in the probability that the positive class is correctly predicted for participants that belong to the positive class, between different groups.

$$\text{EOD} = P(\hat{Y} = 1|Y = 1, A = a) - P(\hat{Y} = 1|Y = 1, A = b), \tag{4.12}$$

where:

- Y is the actual outcome.
- $\hat{Y}$ is the predicted outcome.
- A is the protected attribute (e.g., sex).
- a and b are different groups of the protected attribute (e.g., male and female).

Average Odds Difference

Average Odds Difference (AOD) considers, apart from the true positive (TP) rate, also the false positive (FP) rate. It is computed as the average of the difference of the TP rates and the difference of the FP rates, between different groups.

$$\text{AOD} = \frac{1}{2}\left((FPR_a - FPR_b) + (TPR_a - TPR_b)\right), \tag{4.13}$$

where:

- FPR is the false positive rate, defined as $P(\hat{Y} = 1|Y = 0, A = a)$ for group a.
- TPR is the true positive rate, defined as $P(\hat{Y} = 1|Y = 1, A = a)$ for group a.
- a and b are different groups in the protected attribute.

AOD was chosen to be the main metric, since by considering both the true positive and false positive rates it provides a complete view of the model's fairness.

Fairness Mitigation

Among the fairness mitigation categories, it was chosen to apply mitigation through in-processing, because in this way the model can dynamically learn to make predictions considering both predicted performance and the fairness constraints, rather than being indirectly influenced by the data (pre-processing) or having the decision thresholds adjusted to force that fairness is met (post-processing), which are mainly static methods and do not always preserve data integrity [55]. Specifically, the exponentiated gradient fairness mitigation method was used.

Exponentiated Gradient

Exponentiated Gradient (EG) is an in-processing technique that focuses on the optimization of the parameters considering both the prediction error and the fairness metric [85]. More specifically, this method was used to optimize the complement of the balanced accuracy ($1 -$ balanced accuracy, since EG is designed to minimize the objective function), and the equalized odds difference fairness metric, which considers the differences in TPR and FPR acrosss the groups. At each step, the gradient of the objective function is computed, reflecting how changes in parameters affect both prediction error and fairness. The parameters are updated by multiplying them by the exponential of the negative product of the gradient and

the learning rate (hence the term *exponentiated* in its name), followed by a normalization. The update of the parameters is summarized in equation 4.14.

$$\theta_{\text{new}} = \frac{\theta_{\text{old}} \cdot e^{-\eta \cdot (\nabla L(\theta) + \lambda \nabla F(\theta))}}{\sum \theta_{\text{old}} \cdot e^{-\eta \cdot (\nabla L(\theta) + \lambda \nabla F(\theta))}}, \tag{4.14}$$

where:

- η is the learning rate.
- λ is a regularization parameter to control the importance of the fairness compared to the loss.
- $L(\theta)$ is the gradient of the loss function ($1 -$ balanced accuracy).
- $F(\theta)$ is the gradient of the fairness constraint (equalized odds difference).

Impact of Mitigation on the Prediction

Although Exponentiated Gradient can achieve a good balance between fairness and predictive performance (i.e., balanced accuracy), the latter can be reduced compared to optimizing only for balanced accuracy, as usually happens in a traditional machine learning pipeline. For this reason, it is important to measure the drop in classification performance, so that it can be discussed whether the drop in performance in exchange for the improvement in model fairness can be accepted.

Definition of the privileged and unprivileged groups

Four protected attributes were considered, the fairness of which is important both from the ethical and clinical point of view: sex, age, BMI and cholesterol. Age was categorized into two classes due to the restricted age range of the participants of the UK Biobank (which does not include children and young people): adults and seniors. BMI was categorized into two classes: overweight or obese (being both one class), or otherwise. The variable that represents if a person has high cholesterol was also included for being interesting from a clinical perspective due to its relationship with cardiovascular diseases. Table 4.2 contains the definition of the privileged and unprivileged groups for the protected attributes.

Attribute	Privileged Group	Unprivileged Group
Sex	Males	Females
Cholesterol	No high cholesterol	High cholesterol
Age	Adults: [25, 65)	Seniors: [65, 123)
BMI	Not overweight or obese: [0, 25)	Overweight or obese: [25, ∞)

Table 4.2: Privileged and unprivileged groups for each protected attribute.

Combination of sensitive attributes

In addition to computing the fairness metric for all sensitive attributes, a unique value was computed with a feature that combined them. This feature was calculated as the intersection of the sensitive features, so that the fairness metric is compared between the most possible privileged group (young males not overweight or obese and with no high cholesterol), with respect to the participants that had at least one unprivileged condition. In this way, subjects that belong to one or several unprivileged groups at once are compared with a real privileged group when considering several features, which represents a completely healthy control group in the biomedicine studies analogy.

4.8.4 Feature Importance

Feature importance was examined for the best model using fractals for each disease, which is interesting as they are relatively unexplored features for diagnosing and prognosticating cardiovascular diseases. Feature importance was assessed using a state-of-the art explainability method, namely SHAP [58].

4.9 Tools and Implementation

4.10 Tools

The code for this thesis was developed in Python 3.11.9, and conda 22.11.1 was used to manage the environment and packages. The complete list of the used packages as well as their versions to reproduce the results can be found in annex Annex F.

4.11 Implementation

This thesis required extensive coding, with two Python scripts and three Jupyter notebooks, comprising numerous of cells and lines of code. The description of each code file is detailed below:

- *replication_Pujadas_et_al.py.* It is a Python script to reproduce the results of the table 6 of the supplementary material of the Pujadas et al. (2022) study.
- *diagnosis_prognosis_UKBsample_by_component_outcome.py.* It is a Python script to create the diagnosis and prognosis subcohorts with the data fields of the UK Biobank. It creates files with the positive cases of each disease.
- *performance_evaluation.ipynb.* It is a Jupyter notebook with the whole process for the predictive performance evaluation of all the settings. It also includes the pre-processing of the data, the hyperparameter tuning, and the creation of the final performance tables.
- *fairness_evaluation.ipynb.* It is a Jupyter notebook for the fairness evaluation, including regular fairness results, results after mitigation, and the fairness metrics with a combined feature with all the sensitive attributes.
- *feature_importance.ipynb.* It is a Jupyter notebook for the whole process of the SHAP feature importance computation using the permutation explainer. It also includes the generation and storage of the SHAP summary plots in *.svg* format.

Chapter 5

Experiments and results

5.1 Sample Characteristics

From the 32,003 participants who underwent cardiac magnetic resonance, 72 were removed due to unavailable data for one of the feature sets used, resulting in a total of 31,931 subjects. The prospective mean follow-up period from the imaging visit was 4.96 ($\pm$ 2.24) years. For diagnosis, depending on the disease under study, the number of positive cases was between 210 and 946, being between 234 and 493 in the prognosis scenario. The number of participants in the control group was more than 30,000 in all diseases considered, representing very unbalanced datasets. The specific number of positive and negative cases for each cardiovascular disease is indicated in figure 5.1.

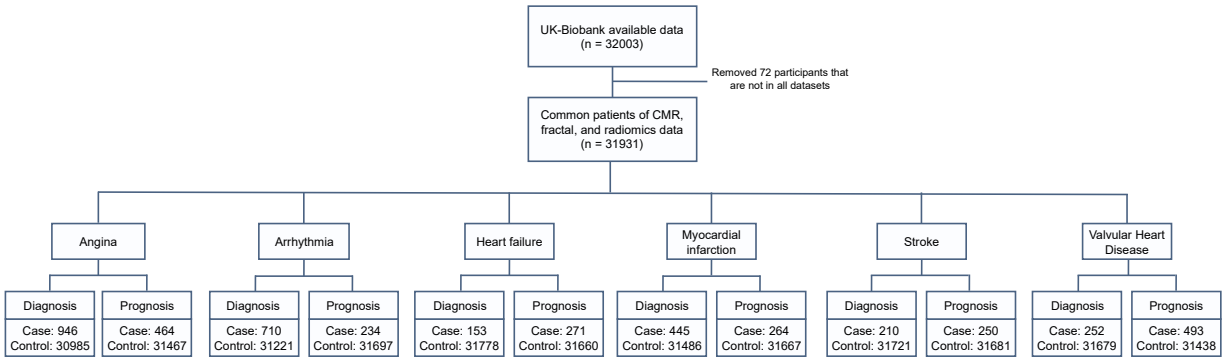


Figure 5.1: Number of cases and controls for diagnosis and prognosis per disease.

Regarding the main characteristics of the study sample, the median age was 64 years, which represents a moderately old population, with all participants between 44 and 82 years old. In terms of BMI, the median value corresponds to the overweight category according to the World Health Organization (WHO) BMI classification. The smoking and diabetes prevalence was relatively small, at 6.3% and 3.1% of the total study sample, respectively, whereas hypertension affected a higher proportion, with 13.7% of the cases. High cholesterol represented the most prevalent risk factor, accounting for 22.6%.

Regarding sex, the number of females and males was almost balanced, with females being 3.6% more of the population than males. The sex differences in the risk factor variables were not very pronounced, except for hypertension and high cholesterol, where the prevalence for females was 6.4% and 5% lower, respectively.

	Female (N=16539)	**Male (N=15392)**	**Overall (N=31931)**
Age			
Mean (SD)	62.6 (7.38)	63.9 (7.64)	63.3 (7.54)
Median [Min, Max]	63.0 [45.0, 82.0]	65.0 [44.0, 82.0]	64.0 [44.0, 82.0]
BMI			
Mean (SD)	26.1 (4.57)	27.1 (3.78)	26.6 (4.24)
Median [Min, Max]	25.2 [15.2, 56.6]	26.7 [16.9, 56.0]	26.0 [15.2, 56.6]
Smoker			
No	15673 (94.8%)	14235 (92.5%)	29908 (93.7%)
Yes	866 (5.2%)	1157 (7.5%)	2023 (6.3%)
Diabetes			
No	16191 (97.9%)	14756 (95.9%)	30947 (96.9%)
Yes	348 (2.1%)	636 (4.1%)	984 (3.1%)
Hypertension			
No	14784 (89.4%)	12782 (83.0%)	27566 (86.3%)
Yes	1755 (10.6%)	2610 (17.0%)	4365 (13.7%)
High cholesterol			
No	13203 (79.8%)	11506 (74.8%)	24709 (77.4%)
Yes	3336 (20.2%)	3886 (25.2%)	7222 (22.6%)

Table 5.1: Sample characteristics by sex and overall.

5.2 Replication of state-of-the-art radiomics results

When replicating the results of Pujadas et al. (2022) [25], the performance metrics achieved were consistent with those reported in their study, suggesting the efficacy of the pipeline. Specifically, focusing on AUC, which was the primary performance metric, there was a small decrease of only 0.02 compared to the results presented by Pujadas et al.

		CMR (Pujadas)	CMR (Colmenero)	Rad (Pujadas)	Rad (Colmenero)	CMR + Rad (Pujadas)	CMR + Rad (Colmenero)
AF	**Accuracy**	0.65 (± 0.04)	0.65 (± 0.02)	0.66 (± 0.05)	0.65 (± 0.03)	0.69 (± 0.04)	0.68 (± 0.02)
	Sensitivity	0.66 (± 0.05)	0.65 (± 0.05)	0.63 (± 0.08)	0.64 (± 0.04)	0.72 (± 0.06)	0.70 (± 0.05)
	Specificity	0.65 (± 0.07)	0.66 (± 0.04)	0.66 (± 0.07)	0.65 (± 0.04)	0.71 (± 0.04)	0.70 (± 0.04)
	AUC	**0.68 (± 0.06)**	**0.68 (± 0.03)**	**0.71 (± 0.08)**	**0.70 (± 0.04)**	**0.71 (± 0.08)**	**0.70 (± 0.04)**
HF	**Accuracy**	0.69 (± 0.03)	0.68 (± 0.03)	0.68 (± 0.05)	0.67 (± 0.03)	0.68 (± 0.06)	0.67 (± 0.01)
	Sensitivity	0.64 (± 0.03)	0.62 (± 0.05)	0.66 (± 0.07)	0.64 (± 0.04)	0.66 (± 0.04)	0.64 (± 0.06)
	Specificity	0.74 (± 0.08)	0.72 (± 0.05)	0.70 (± 0.07)	0.68 (± 0.04)	0.71 (± 0.09)	0.69 (± 0.04)
	AUC	**0.74 (± 0.03)**	**0.73 (± 0.02)**	**0.77 (± 0.05)**	**0.75 (± 0.03)**	**0.76 (± 0.5)**	**0.74 (± 0.03)**
MI	**Accuracy**	0.63 (± 0.03)	0.63 (± 0.05)	0.64 (± 0.05)	0.63 (± 0.03)	0.64 (± 0.05)	0.63 (± 0.04)
	Sensitivity	0.62 (± 0.09)	0.62 (± 0.05)	0.64 (± 0.01)	0.63 (± 0.03)	0.64 (± 0.09)	0.63 (± 0.06)
	Specificity	0.63 (± 0.04)	0.61 (± 0.06)	0.64 (± 0.05)	0.62 (± 0.04)	0.64 (± 0.05)	0.62 (± 0.05)
	AUC	**0.68 (± 0.04)**	**0.67 (± 0.05)**	**0.69 (± 0.06)**	**0.67 (± 0.03)**	**0.68 (± 0.06)**	**0.66 (± 0.05)**

Table 5.2: Performance replication of table 6 of the supplementary material of the Pujadas et al. (2022) study, where the controls represent people completely free from cardiovascular disease.

5.3 Predictive Performance

Given the extensive testing of various settings, including different feature combinations and diseases for diagnosis and prognosis, the primary analysis will concentrate on the most common diseases, angina and arrhythmia, to prevent an overload of information for the reader. In addition, a summary of the results for other diseases will be also analyzed, to also provide the most important findings for the other conditions. The complete performance tables for each condition can be found in annex Annex C.

5.3.1 Diagnosis

Angina

The predictive performance results for angina diagnosis show values of the mean balanced accuracy across the 3 folds used for testing that ranges from 0.617 to 0.665. When comparing models that employ fractal-based features to the models using radiomics-based features, they achieved similar performance results, being the ones for radiomics slightly higher (maximum 0.013 difference). CMR models also obtained a similar performance compared to the ones for radiomics and fractals, being slightly higher or lower depending on the model. The best overall performance was obtained by the XGBoost model with the feature set that combines both radiomics and fractal features, that yielded a mean balanced accuracy of 0.665. In contrast, SVM got the lowest performance in all the scenarios (regardless of the feature set).

	CMR	**Rad**	**Frac**	**CMR+Rad**	**CMR+Frac**	**Rad+Frac**	**All**
XGB	0.644 ± 0.03	0.657 ± 0.02	0.647 ± 0.01	0.655 ± 0.02	0.644 ± 0.02	**0.665 ± 0.01**	0.655 ± 0.02
Ada	0.636 ± 0.02	0.648 ± 0.01	0.651 ± 0.01	**0.658 ± 0.01**	0.65 ± 0.03	0.652 ± 0.01	0.653 ± 0.01
LGBM	0.65 ± 0.02	0.656 ± 0.02	0.643 ± 0.01	0.652 ± 0.02	0.645 ± 0.02	**0.659 ± 0.01**	0.656 ± 0.02
RF	0.648 ± 0.02	0.647 ± 0.01	0.644 ± 0.02	0.643 ± 0.02	**0.651 ± 0.02**	0.644 ± 0.01	0.645 ± 0.02
BRF	0.649 ± 0.03	0.655 ± 0.01	0.648 ± 0.02	0.658 ± 0.02	**0.662 ± 0.03**	0.653 ± 0.02	0.655 ± 0.03
SVM	0.617 ± 0.03	**0.636 ± 0.01**	0.624 ± 0.02	0.633 ± 0.02	0.621 ± 0.03	0.635 ± 0.02	0.634 ± 0.02

Table 5.3: Performance of all models and feature combinations for angina diagnosis. Performance is measured by the balanced accuracy across the outer (test) CV folds, indicating the standard deviation as well. For each model, the result that represents the best performance is highlighted in bold. The overall higher performance result is highlighted in blue. 'Rad' and 'Frac' are acronyms for Radiomics and Fractal, respectively.

Arrhythmia

Regarding arrhythmia diagnosis, a slightly better performance was obtained compared to angina, with mean balanced accuracy values ranging from 0.620 to 0.678. Similarly, the predictive performance of the radiomics and fractals models were alike. The best overall performance was obtained by the LightGBM model using all the feature sets (CMR, radiomics and fractals), that yielded a mean balanced accuracy of 0.678.

	CMR	Rad	Frac	CMR+Rad	CMR+Frac	Rad+Frac	All
XGB	0.653 ±0.01	0.664 ±0.01	0.638 ±0.02	0.668 ±0.01	0.646 ±0.01	0.660 ±0.00	**0.672 ± 0.00**
Ada	0.639 ±0.01	0.655 ±0.02	0.635 ±0.00	0.659 ±0.01	0.645 ±0.01	**0.666 ± 0.02**	0.664 ±0.00
LGBM	0.645 ±0.00	0.670 ±0.01	0.631 ±0.02	0.665 ±0.01	0.655 ±0.01	0.661 ±0.02	**0.678 ± 0.01**
RF	0.642 ±0.00	0.647 ±0.03	0.637 ±0.02	0.660 ±0.00	0.655 ±0.02	**0.661 ± 0.01**	0.660 ±0.01
BRF	0.647 ±0.01	0.656 ±0.00	0.639 ±0.01	0.656 ±0.00	0.652 ±0.01	0.658 ±0.01	**0.662 ± 0.01**
SVM	0.622 ±0.01	0.641 ±0.01	0.620 ±0.01	0.647 ±0.00	0.634 ±0.01	0.649 ±0.01	**0.654 ± 0.01**

Table 5.4: Performance of all models and feature combinations for arrhythmia diagnosis.

Summary of the diagnosis results

When comparing the mean balanced accuracy of the best fractal and radiomics models, the one for radiomics is slightly better. However, in the Wilcoxon Signed-Rank test, statistical significance was not reached (a p-value >0.05 was observed for all diseases), indicated by the absence of an asterisk symbol in the table 5.5. Thus, the null hypothesis is accepted and it is considered that there are no significant differences between the performance of both models. In addition, the fractal feature set was the only one that was included in all the best overall models for all the diseases, either along with CMR indices, Radiomics features, or both.

Regarding the best overall models, the highest mean balanced accuracy values were obtained for heart failure and myocardial infarction, achieving 0.748 and 0.715 scores, respectively. Valvular heart disease got a similar performance to that of angina and arrhythmia, with 0.672 balanced accuracy. The worst predictive performance was obtained for stroke, showing a 0.608 score. Moreover, there was not a general pattern for a model to stand out from the rest, but LightGBM, which yielded the highest performance for three diseases.

	Best radiomics model		Best fractal model		Best overall model		
	Model	**Perf.**	**Model**	**Perf.**	**Model**	**Features**	**Perf.**
Angina	XGB	0.657	Ada	0.651	XGBoost	Rad + Frac	0.665
Arrhythmia	LGBM	0.67	BRF	0.639	LGBM	All	0.678
Heart Failure	RF	0.739	BRF	0.707	LGBM	All	0.748
MI	Ada	0.707	XGB	0.704	LGBM	All	0.715
Stroke	RF	0.597	XGB	0.589	BRF	Rad + Frac	0.608
VHD	RF	0.656	XGB	0.652	BRF	CMR + Frac	0.672

Table 5.5: Mean outer CV (testing) balanced accuracy scores of the best radiomics and fractal models, as well as the best overall model, for all component outcomes in the diagnosis scenario. If there is a statistically significant difference (p-value < 0.05) between the performance of the best fractal and radiomics models in the Wilcoxon Signed-Rank test, it is indicated with an asterisk.

5.3.2 Prognosis

Angina

In the performance results for angina prognosis, the mean balanced accuracy scores ranged from 0.580 to 0.635. The performance of the models with CMR indices, radiomics and fractal features were similar, although the one for radiomics was slightly lower. The model with the better performance was the balanced random forest with fractal features, which achieved a 0.635 score, although the difference with other models was not very pronounced.

	CMR	**Rad**	**Frac**	**CMR+Rad**	**CMR+Frac**	**Rad+Frac**	**All**
XGB	**0.624 ± 0.01**	0.608 ±0.01	0.622 ±0.01	0.613 ±0.04	0.618 ±0.02	0.615 ±0.01	0.605 ±0.03
Ada	**0.624 ± 0.01**	0.620 ±0.03	0.614 ±0.01	0.603 ±0.03	0.617 ±0.01	0.612 ±0.01	0.602 ±0.02
LGBM	0.609 ±0.03	0.618 ±0.02	**0.629 ± 0.02**	0.600 ±0.03	0.601 ±0.02	0.624 ±0.02	0.601 ±0.03
RF	**0.623 ± 0.01**	0.616 ±0.01	0.621 ±0.00	0.615 ±0.03	0.620 ±0.01	0.608 ±0.02	0.606 ±0.03
BRF	0.634 ±0.01	0.618 ±0.01	**0.635 ± 0.00**	0.617 ±0.02	0.630 ±0.02	0.614 ±0.03	0.615 ±0.01
SVM	**0.599 ± 0.03**	0.598 ±0.02	0.598 ±0.01	0.590 ±0.03	0.580 ±0.02	0.594 ±0.01	0.592 ±0.02

Table 5.6: Performance of all models and feature combinations for angina prognosis.

Arrhythmia

Regarding arrhythmia, the mean balanced accuracy scores ranged from 0.552 to 0.629. In general, the performance of the CMR indices models was superior to those that used radiomics or fractals, especially when using random forest-based models. Radiomics and fractal performance was similar, although fractal models achieved a slightly higher men balanced accuracy for all but one model. The best performance was obtained by the balanced random forest with the CMR indices, that yielded a 0.629 balanced accuracy.

	CMR	**Rad**	**Frac**	**CMR+Rad**	**CMR+Frac**	**Rad+Frac**	**All**
XGB	0.606 ±0.01	**0.607 ± 0.02**	0.604 ±0.04	0.578 ±0.01	0.600 ±0.01	0.598 ±0.02	0.578 ±0.03
Ada	0.599 ±0.01	0.577 ±0.03	**0.601 ± 0.01**	0.570 ±0.01	0.581 ±0.03	0.574 ±0.02	0.571 ±0.02
LGBM	**0.608 ± 0.01**	0.572 ±0.01	**0.608 ± 0.01**	0.581 ±0.03	0.585 ±0.02	0.581 ±0.02	0.580 ±0.03
RF	**0.620 ± 0.02**	0.580 ±0.05	0.599 ±0.01	0.598 ±0.01	0.600 ±0.02	0.584 ±0.03	0.593 ±0.01
BRF	**0.629 ± 0.01**	0.590 ±0.03	0.613 ±0.01	0.605 ±0.02	0.613 ±0.02	0.593 ±0.03	0.612 ±0.03
SVM	**0.561 ± 0.01**	0.552 ±0.03	0.557 ±0.01	0.559 ±0.03	**0.561 ± 0.02**	0.555 ±0.02	0.560 ±0.01

Table 5.7: Performance of all models and feature combinations for arrhythmia prognosis.

Summary of the prognosis results

Wilcoxon Signed-Rank test, showed no statistically significant differences between the best fractal and radiomics models for any of the diseases, despite being the performance for the fractal models slightly higher in general terms. In addition, fractal features were included in the feature set of the best overall model for four diseases, whereas radiomics features were not included in none any of them.

Regarding the best overall performance results, the one for heart failure achieved the highest predictive performance, with a 0.732 balanced accuracy score, followed by valvular heart disease (0.698). In contrast, the lowest overall results were for angina and arrhythmia, with 0.635 and 0.629, respectively. Random forest-based models led the best results, especially balanced random forest, which was the best model for four diseases.

	Best radiomics model		Best fractal model		Best overall model		
	Model	**Perf.**	**Model**	**Perf.**	**Model**	**Features**	**Perf.**
Angina	Ada	0.62	BRF	0.635	BRF	Frac	0.635
Arrhythmia	XGBoost	0.607	BRF	0.613	BRF	CMR	0.629
Heart Failure	Ada	0.725	BRF	0.69	BRF	CMR + Frac	0.732
MI	XGBoost	0.612	Ada	0.64	BRF	CMR + Frac	0.661
Stroke	BRF	0.602	Ada	0.619	RF	CMR + Frac	0.641
VHD	LGBM	0.681	XGB	0.685	RF	CMR	0.698

Table 5.8: Mean outer CV (testing) balanced accuracy scores of the best radiomics and fractal models, as well as the best overall model, for all component outcomes in the prognosis scenario. If there is a statistically significant difference (p-value < 0.05) between the performance of the best fractal and radiomics models in the Wilcoxon Signed-Rank test, it is indicated with an asterisk.

5.4 Fairness

The difference of all the used fairness metrics (demographic parity difference, equal opportunity difference, and average odds difference) was computed as:

$$\text{Fairness Difference} = \text{Metric}_{\text{unprivileged}} - \text{Metric}_{\text{privileged}} \tag{5.1}$$

Thus, values closer to 1 favours the unprivileged group, whereas values closer to -1 favours the privileged group. Although the most desirable situation is that there is no bias (the fairness metric is 0), when bias is inevitable, it is considered preferable to have bias toward the unprivileged group compared to having it toward the privileged one.

Similar to the predictive performance results, due to the large amount of information, some results will be analyzed in depth, while a summary will be provided and analyzed for others. Specifically, the focus will be on angina fairness, both with and without mitigation, for diagnosis and prognosis settings. In addition, for all diseases, results will be shown about the feature combining all sensitive attributes. The complete tables with the fairness metrics for all diseases can be found in Annex D.

5.4.1 Diagnosis

Angina

In the angina fairness results for diagnosis, the sensitive attributes with a greater bias were age and sex, which showed fairness scores between 0.49 and 0.78 for age, and between -0.33 and -0.44 for sex, indicating a large deviation from the ideal case where fairness metrics are 0. In these sensitive attributes, the demographic parity (DP) difference suggested a great imbalance in the probability of receiving the positive outcome, with older and male participants (separately) being more likely to be predicted with an angina diagnosis, since $DP_{senior} - DP_{adult} > 0$ and $DP_{female} - DP_{male} < 0$, respectively.

The differences between the privileged and unprivileged groups of sex and age were also large in relation to the equal opportunity difference (EOD), with the true positive rate (TPR) for seniors, and the one for males, being superior compared to the other group. Thus, the probability that the model correctly identified a diagnosis among all actual diagnosis cases was considerably greater in those groups.

Similarly, the average odds difference (AOD) also showed notable differences between the sex and age privileged and unprivileged groups. The disparity in the average of the true positive rate (TPR) and false positive rate (FPR) differences was then higher in those groups. Then, even when considering the FPR as well, there were still remarkable differences.

In contrast, there was a lower bias in the hypertension and overweight or obese sensitive attributes, particularly the latter, which ranged from 0.07 to 0.22.

Regarding the best fractal and radiomics models comparison, the radiomics model was fairer in sex and age for all the fairness metrics, whereas the fractal model was fairer in hypertension and being overweight or obese.

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	-0.439 ± (0.10)	**0.273 ± (0.01)**	**0.106 ± (0.01)**	0.778 ± (0.02)
Eq. Opp. Diff. (Rad)	**-0.384 ± (0.07)**	0.291 ± (0.00)	0.223 ± (0.02)	**0.552 ± (0.08)**
Avg. Odds. Diff. (Frac)	-0.409 ± (0.08)	**0.187 ± (0.01)**	**0.071 ± (0.01)**	-0.739 ± (0.03)
Avg. Odds. Diff. (Rad)	**-0.346 ± (0.06)**	0.240 ± (0.02)	0.200 ± (0.02)	**0.516 ± (0.06)**
Dem. Parity Diff. (Frac)	-0.388 ± (0.06)	**0.278 ± (0.01)**	**0.111 ± (0.01)**	0.705 ± (0.04)
Dem. Parity Diff. (Rad)	**-0.332 ± (0.03)**	0.303 ± (0.01)	0.217 ± (0.00)	**0.489 ± (0.04)**

Table 5.9: Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) for angina diagnosis. The fairness metrics include the equal opportunity difference, the average odds difference, and the demographic parity difference. For every metric and sensitive attribute, the fairest result between the radiomics and fractal models is highlighted in bold.

When applying mitigation though the exponentiated gradient in-processing technique, all fairness metrics scores were improved, as indicated by values closer to 0. The disparity between the sensitive attributes that suffered from a greater biased (age and sex) was largely reduced, showing decreases between 0.2 and 0.7. However, sensitive attributes that initially had a lower bias (hypertension and overweights or obese) showed a smaller improvement, with hypertension only moving between 0.002 and 0.08 units closer to 0.

After mitigation, the best fractal model achieved the best performance for all sensitive attributes, with fairness scores up to 0.05 for all of them but for hypertension. Regarding the balanced accuracy drop due to applying mitigation, radiomics performance was reduced to a slightly lesser extent, although the difference was not very pronounced.

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	**-0.036 ± (0.01)**	**0.196 ± (0.05)**	**0.091 ± (0.04)**	**0.051 ± (0.01)**
Eq. Opp. Diff. (Rad)	-0.172 ± (0.06)	0.259 ± (0.01)	0.157 ± (0.01)	0.181 ± (0.02)
Avg. Odds. Diff. (Frac)	**-0.025 ± (0.01)**	**0.138 ± (0.02)**	**0.014 ± (0.04)**	**0.015 ± (0.03)**
Avg. Odds. Diff. (Rad)	-0.119 ± (0.05)	0.209 ± (0.02)	0.129 ± (0.04)	0.147 ± (0.02)
Dem. Parity Diff. (Frac)	**-0.042 ± (0.01)**	**0.183 ± (0.07)**	**0.069 ± (0.01)**	**0.034 ± (0.02)**
Dem. Parity Diff. (Rad)	-0.099 ± (0.02)	0.270 ± (0.01)	0.162 ± (0.01)	0.147 ± (0.03)
Bal. Acc. Drop (Frac)	0.027	0.052	0.024	0.063
Bal. Acc. Drop (Rad)	**0.022**	**0.011**	**0.011**	**0.027**

Table 5.10: Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) when applying mitigation (through exponentiated gradient) for angina diagnosis. The drop in the balanced accuracy performance is indicated as well.

5.4.2 Combination of sensitive attributes

In the fairness results of the feature that combines all the sensitive attributes through their intersection, all the fairness metric scores where positive, meaning that the privileged group (non-hypertensive and non-overweight young males) was more favoured compared to the unprivileged one.

In general, angina prediction suffered the greatest bias, with AOD values between 0.495 and 0.510. Focusing on AOD, which considers both the TPR and the FPR, radiomics models were superior to those for fractals for angina, arrhythmia, heart failure, and stroke, whereas fractal models were better for myocardial infarction and valvular heart disease.

	Angina	Arrhythmia	HF	MI	Stroke	VHD
EOD (Frac)	**0.495 ± (0.12)**	0.233 ± (0.03)	0.222 ± (0.07)	**0.156 ± (0.06)**	0.524 ± (0.20)	0.184 ± (0.1)
EOD (Rad)	0.510 ± (0.14)	**0.216 ± (0.21)**	**0.217 ± (0.06)**	0.372 ± (0.3)	**0.290 ± (0.09)**	**0.159 ± (0.15)**
AOD (Frac)	0.403 ± (0.07)	0.135 ± (0.02)	0.017 ± (0.14)	**0.009 ± (0.11)**	0.360 ± (0.13)	**0.009 ± (0.1)**
AOD (Rad)	**0.387 ± (0.08)**	**0.115 ± (0.12)**	**0.015 ± (0.11)**	0.182 ± (0.14)	**0.024 ± (0.10)**	0.074 ± (0.1)
DP (Frac)	0.320 ± (0.01)	0.043 ± (0.01)	**0.056 ± (0.04)**	0.049 ± (0.03)	0.197 ± (0.12)	0.112 ± (0.05)
DP (Rad)	**0.274 ± (0.03)**	**0.041 ± (0.04)**	0.076 ± (0.07)	**0.006 ± (0.00)**	**0.123 ± (0.03)**	**0.054 ± (0.02)**

Table 5.11: Fairness results with the combined sensitive attribute in the diagnosis scenario. The used acronyms for the fairness metrics are equal opportunity difference (EOD), average odds difference (AOD), demographic parity difference (DPD, and balanced accuracy drop (BAD). For the main outcomes, the used acronyms are heart failure (HF), myocardial infarction (MI), and valvular heart disease (VHD).

When mitigation was applied, every fairness metric was improved. Regarding AOD, fractal models were superior in all diseases but stroke. Mitigation did not lead to a big drop in the predictive performance, and the differences between the fractal and radiomics models were not very profound except for arrhythmia and heart failure, where the performance decrease of the radiomics model was between 0.07 and 0.09 lower.

	Angina	Arrhythmia	HF	MI	Stroke	VHD
EOD (Frac)	0.294 ± (0.10)	**0.101 ± (0.11)**	**0.176 ± (0.10)**	**0.150 ± (0.12)**	0.518 ± (0.06)	0.181 ± (0.16)
EOD (Rad)	**0.239 ± (0.07)**	0.207 ± (0.2)	0.211 ± (0.06)	0.367 ± (0.29)	**0.285 ± (0.09)**	**0.099 ± (0.08)**
AOD (Frac)	**0.174 ± (0.06)**	**0.030 ± (0.06)**	**0.003 ± (0.1)**	**0.006 ± (0.12)**	0.353 ± (0.05)	**0.005 ± (0.18)**
AOD (Rad)	0.186 ± (0.03)	0.112 ± (0.12)	0.012 ± (0.11)	0.174 ± (0.13)	**0.021 ± (0.11)**	0.041 ± (0.06)
DPD (Frac)	**0.062 ± (0.05)**	**0.011 ± (0.01)**	**0.004 ± (0.00)**	0.039 ± (0.03)	0.133 ± (0.08)	0.103 ± (0.05)
DPD (Rad)	0.162 ± (0.03)	0.039 ± (0.02)	0.067 ± (0.07)	**0.004 ± (0.02)**	**0.121 ± (0.03)**	**0.045 ± (0.02)**
BA Drop (Frac)	0.018	0.073	0.089	**0.001**	**0.004**	**0.004**
BA Drop (Rad)	**0.002**	**0.003**	**0.001**	0.002	0.005	**0.004**

Table 5.12: Fairness results with the combined sensitive attribute in the diagnosis scenario when applying mitigation (through exponentiated gradient).

5.4.3 Prognosis

Angina

In the fairness metrics for angina prognosis, the greatest disparities were also found for age and sex, with males and adults being favoured in their probability of being predicted as a diagnosis case, their TPR and, both their TPR and FPR, as indicated by the DP diffence, EOD, and AOD metrics, respectively.

Similar to the diagnosis scenario, the sensitive attributes corresponding to overweight or obese was the one with the lower disparity. Focusing on AOD, the best fractal model was fairer for hypertension and overweight or obese, whereas the best radiomics model was the best for sex and age.

	Sex	**Hypertension**	**Overweight or obese**	**Age**
Eq. Opp. Diff. (Frac)	-0.594 ± (0.06)	0.271 ± (0.02)	**0.141 ± (0.02)**	0.674 ± (0.09)
Eq. Opp. Diff. (Rad)	**-0.337 ± (0.10)**	**0.255 ± (0.03)**	0.185 ± (0.02)	**0.493 ± (0.16)**
Avg. Odds. Diff. (Frac)	-0.551 ± (0.04)	**0.200 ± (0.03)**	**0.098 ± (0.02)**	0.633 ± (0.07)
Avg. Odds. Diff. (Rad)	**-0.301 ± (0.09)**	0.212 ± (0.03)	0.156 ± (0.02)	**0.446 ± (0.15)**
Dem. Parity Diff. (Frac)	-0.514 ± (0.03)	0.279 ± (0.02)	**0.145 ± (0.02)**	0.598 ± (0.06)
Dem. Parity Diff. (Rad)	**-0.283 ± (0.07)**	**0.267 ± (0.04)**	0.189 ± (0.02)	**0.427 ± (0.11)**

Table 5.13: Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) for angina prognosis.

After mitigation, all the fairness metrics were improved for every sensitive attribute and model. In this case, the best fractal model was superior to the best radiomics model in all the settings. However, mitigation affected less severely to the radiomics model, with differences in the balanced accuracy drop up to 0.12.

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	**-0.044 ± (0.06)**	**0.006 ± (0.01)**	**0.020 ± (0.02)**	**0.112 ± (0.04)**
Eq. Opp. Diff. (Rad)	-0.144 ± (0.05)	0.223 ± (0.01)	0.112 ± (0.05)	0.174 ± (0.07)
Avg. Odds. Diff. (Frac)	**-0.023 ± (0.04)**	**0.002 ± (0.00)**	**0.013 ± (0.02)**	**0.059 ± (0.03)**
Avg. Odds. Diff. (Rad)	-0.115 ± (0.05)	0.172 ± (0.01)	0.098 ± (0.04)	0.133 ± (0.08)
Dem. Parity Diff. (Frac)	**-0.010 ± (0.01)**	**0.001 ± (0.00)**	**0.011 ± (0.01)**	**0.018 ± (0.01)**
Dem. Parity Diff. (Rad)	-0.102 ± (0.07)	0.230 ± (0.01)	0.117 ± (0.04)	0.120 ± (0.11)
Bal. Acc. Drop (Frac)	0.131	0.142	0.128	0.122
Bal. Acc. Drop (Rad)	**0.018**	**0.023**	**0.012**	**0.041**

Table 5.14: Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) when applying mitigation (through exponentiated gradient) for angina prognosis.

5.4.4 Combination of sensitive attributes

The fairness results of the feature that combines all the defined sensitive attributes showed that angina's prediction suffered the greatest bias, similar to the diagnosis results. Also, focusing on AOD, all diseases but arrhythmia and myocardial infarction showed a lower disparity with the radiomics models.

	Angina	Arrhythmia	HF	MI	Stroke	VHD
EOD (Frac)	0.517 ± (0.15)	0.215 ± (0.16)	0.382 ± (0.23)	**0.184 ± (0.11)**	0.436 ± (0.37)	0.287 ± (0.17)
EOD (Rad)	**0.491 ± (0.25)**	**0.181 ± (0.19)**	**0.138 ± (0.07)**	0.312 ± (0.29)	**0.250 ± (0.13)**	**0.186 ± (0.17)**
AOD (Frac)	0.387 ± (0.09)	**0.053 ± (0.11)**	0.114 ± (0.23)	**0.028 ± (0.14)**	0.290 ± (0.21)	0.071 ± (0.17)
AOD (Rad)	**0.302 ± (0.21)**	0.081 ± (0.14)	**0.045 ± (0.06)**	0.144 ± (0.13)	**0.003 ± (0.13)**	**0.061 ± (0.11)**
DP (Frac)	0.267 ± (0.05)	0.055 ± (0.01)	**0.068 ± (0.03)**	0.099 ± (0.05)	0.145 ± (0.06)	0.107 ± (0.03)
DP (Rad)	**0.215 ± (0.05)**	**0.052 ± (0.05)**	0.074 ± (0.03)	**0.036 ± (0.02)**	**0.030 ± (0.01)**	**0.032 ± (0.02)**

Table 5.15: Fairness results with the combined sensitive attribute in the prognosis scenario.

Upon implementing mitigation, the fairness metrics scores were all improved, in line with previous results. Regarding AOD, fractal models were fairer compared to radiomics models for all diseases except heart failure and stroke. The mean balanced accuracy decreased to a lower extent with the radiomics models, although the difference in drop was only relevant for angina, where radiomics performance decreased by 0.1 less.

	Angina	**Arrhythmia**	**HF**	**MI**	**Stroke**	**VHD**
EOD (Frac)	**0.043 ± (0.04)**	**0.075 ± (0.11)**	0.374 ± (0.21)	**0.134 ± (0.08)**	0.433 ± (0.29)	0.283 ± (0.02)
EOD (Rad)	0.292 ± (0.11)	0.155 ± (0.13)	**0.131 ± (0.06)**	0.307 ± (0.26)	**0.131 ± (0.15)**	**0.184 ± (0.15)**
AOD (Frac)	**0.016 ± (0.02)**	**0.035 ± (0.05)**	0.108 ± (0.11)	**0.011 ± (0.10)**	0.280 ± (0.13)	**0.001 ± (0.17)**
AOD (Rad)	0.155 ± (0.11)	0.075 ± (0.1)	**0.040 ± (0.06)**	0.112 ± (0.18)	**0.001 ± (0.07)**	0.056 ± (0.09)
DP (Frac)	**0.012 ± (0.01)**	**0.002 ± (0.00)**	**0.022 ± (0.01)**	0.063 ± (0.00)	0.079 ± (0.04)	0.096 ± (0.03)
DP (Rad)	0.119 ± (0.02)	0.045 ± (0.03)	0.070 ± (0.03)	**0.031 ± (0.01)**	**0.011 ± (0.01)**	**0.030 ± (0.02)**
BAD (Frac)	0.123	0.096	0.005	0.008	0.029	0.003
BAD (Rad)	**0.014**	**0.001**	**0.001**	**0.001**	**0.025**	**0.002**

Table 5.16: Fairness results with the combined sensitive attribute in the prognosis scenario when applying mitigation (through exponentiated gradient).

5.5 Feature Importance

Similarly to previous sections, the feature importance results of angina and arrhythmia will be analyzed in more detail, while analyzing the results of the rest of the diseases in a broader context through their 3 most important features. The disease-specific SHAP summary plots for the other conditions can be found in Annex E.

5.5.1 Diagnosis

Angina

The SHAP values of the best fractal model for angina diagnosis show that the feature that affects the model output the most is age, with a big difference compared to the rest. The second most important feature is sex, followed by the features about the fractal dimension (both for end-diastole and end-systole), the lacunarity of right (ED) and left (ES) ventricle, and the fractal dimensions of the left ventricle (both ED and ES).

When analyzing the direction of the effect, it is observed that being older, being male, as well as having high values of the myocardium fractal dimensions and the right ventricle lacunarity (ED), shift the model's predicted probability towards angina diagnosis. In contrast, higher values of the lacunarity and fractal dimensions of the left ventricle shift the model's prediction towards a non-angina diagnosis.

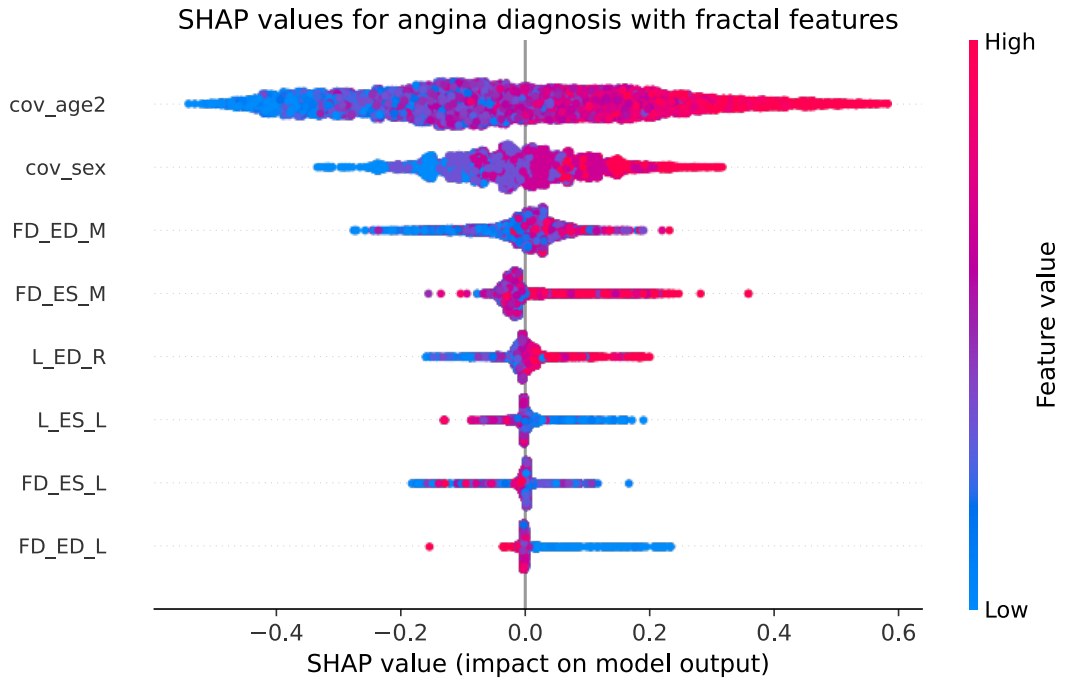


Figure 5.2: SHAP values of the top 8 feature of the best model for angina diagnosis using fractal-based features.

Arrhythmia

Similar to angina, the most important feature for arrhythmia diagnosis is age. The next most relevant features are the myocardium fractal dimension and lacunarity (both for ES), the left ventricle fractal dimension (ES and ED), sex, and the left and right ventricles lacunarity (both for ES).

Being male, and having higher values of age, myocardium and right ventricle lacunarity (both for ES), and left ventricle fractal dimension (ES), increase the model's prediction toward an arrhythmia diagnosis. Conversely, higher values of the myocardium fractal dimension (ES), and the left and right ventricle lacunarity (both for ES) decrease the model's predicted probability toward a non-arrhythmia diagnosis.

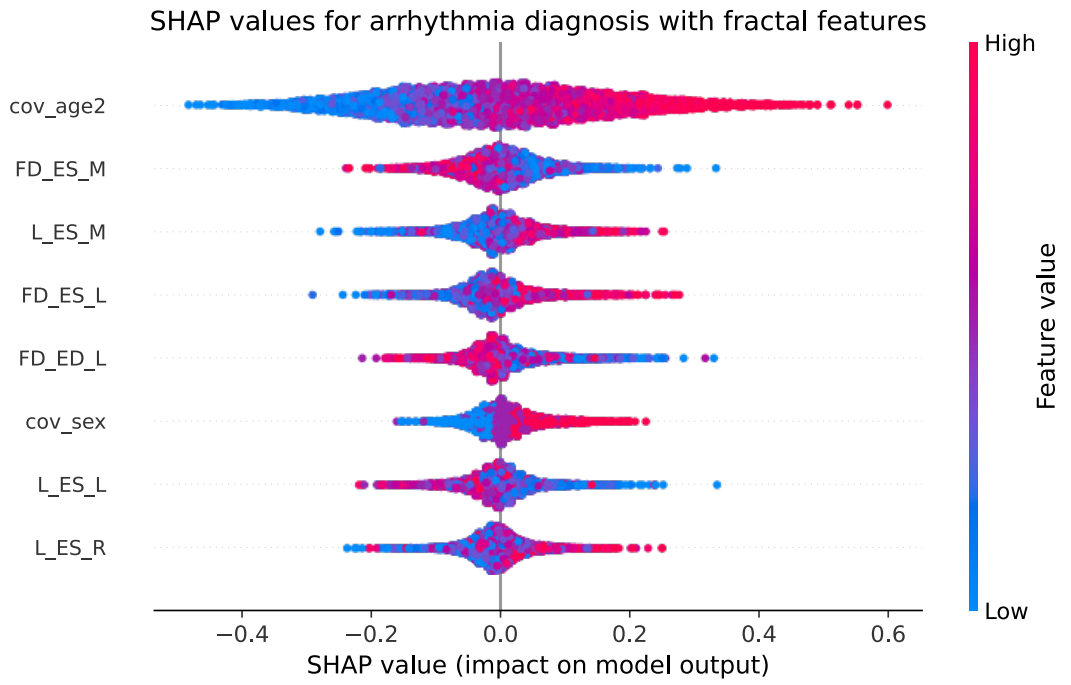


Figure 5.3: SHAP values of the top 8 feature of the best model for arrhythmia diagnosis using fractal-based features.

5.5.2 Summary of all the feature importance of all diseases

Regarding the three most important features based on the SHAP values magnitude for diagnosis, age was in all the top 3 features of the different diseases, and sex was in all of them but valvular heart disease and heart failure. Regarding the fractal features, the features of the myocardium were generally the most important ones, being in all the top-3 features of the diseases except for stroke. The fractal features for the left ventricle were important for stroke and valvular heart disease.

Condition	Top 3 Features
Angina	Age, sex, MYO (FD)
Arrhythmia	Age, MYO (FD), MYO (L)
Heart failure	MYO (FD), age, MYO (L)
MI	sex, age, MYO (L)
Stroke	age, sex, LV (L)
VHD	age, MYO (L), LV (L)

Table 5.17: Top 3 features based on SHAP values for every disease in the diagnosis scenario. FD and L represent fractal dimension and lacunarity, respectively.

5.5.3 Prognosis

Angina

Similar to angina diagnosis, the most relevant feature according to the SHAP values is age. The next more important features are sex, the myocardial fractal dimension (ED and ES), the myocardium and right ventricle lacunarity (ED), and the myocardium and left ventricle lacunarity (ES).

Considering the sign of the SHAP values, it is observed that higher values of all the mentioned features shift the model's predicted probability toward angina prognosis, but the feature representing the myocardium lacunarity (ED), where higher values shift the model's prediction toward a non-angina prognosis.

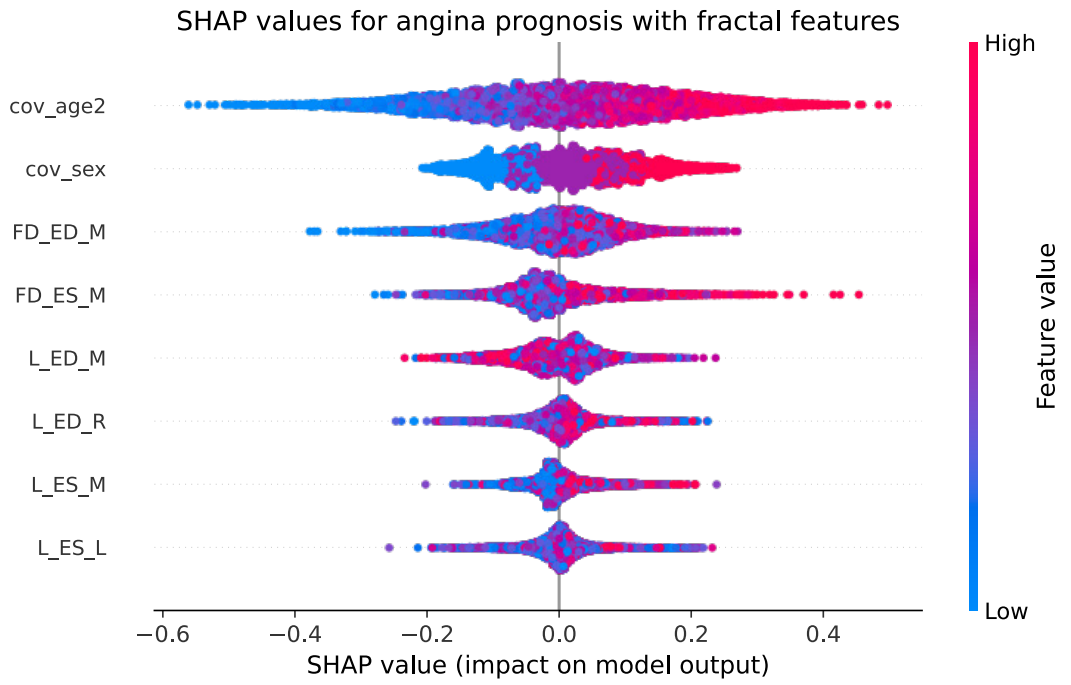


Figure 5.4: SHAP values of the top 8 feature of the best model for angina prognosis using fractal-based features.

Arrhythmia

For arrhythmia diagnosis, the most important feature is once again age. In this case, the difference between the first and second most relevant features is not very pronounced, with the contribution of the left ventricle fractal dimension (ES) being considerable as well. In the feature importance rank, this feature is followed by left and right ventricle lacunarity (both for ES), myocardium fractal dimension (ED), left ventricle and myocardium lacunarity, and sex.

Being older, being male, and higher left (ED) and right (ES) ventricle lacunarity increase the model's prediction toward a higher probability of arrhythmia prognosis. In contrast, higher values of myocardium (ED) and left ventricle (ES) lacunarity decrease the model's predicted probability toward a non-arrhythmia prognosis.

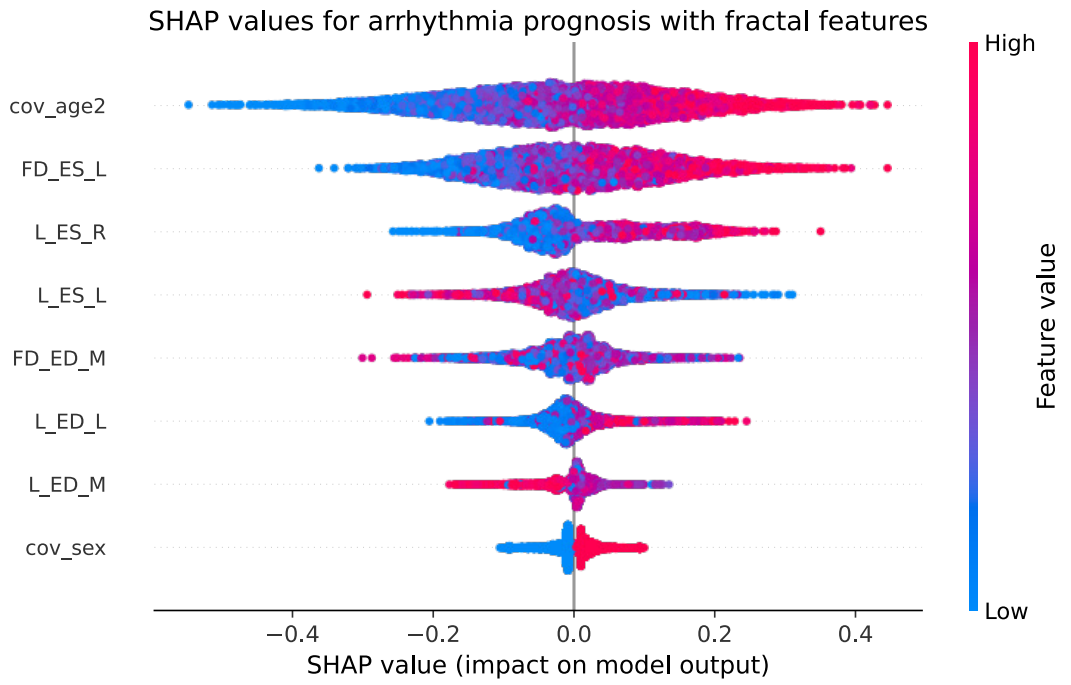


Figure 5.5: SHAP values of the top 8 feature of the best model for arrhythmia prognosis using fractal-based features.

5.5.4 Summary of all the feature importance of all diseases

Regarding the three most important features for prognosis, age was among them for all the diseases. The myocardium fractal features were also very relevant for three diseases, and the ones for the left and right ventricle were in the top 3 features for two conditions each.

Condition	Top 3 Features
Angina	Age, sex, MYO (FD)
Arrhythmia	Age, LV (FD), RV (L)
Heart failure	Age, LV (FD), LV (L)
MI	Sex, age, MYO (L)
Stroke	Age, RV (FD) both ES and ED
VHD	Age, MYO (FD), MYO (L)

Table 5.18: Top 3 features based on SHAP values for every disease in the prognosis scenario.

Chapter 6

Discussion and conclusions

6.1 Discussion

6.1.1 Discussion of the results

Replication of state-of-the-art radiomics results

When replicating the results of Pujadas et al. (2022) [25], there was a slight decrease in the performance, which could be explained by the different method used for feature selection, which was based on the weight magnitude, whereas in the original study they used sequential forward feature selection (SFFS). Unlike focusing on indirect feature importance through model weights, SFFS directly focuses on performance improvement and is a more exhaustive approach that iteratively evaluates each feature at a time, which could better capture the synergistic effects of features, potentially leading to higher performance. Nevertheless, despite not using SFFS, the obtained performance metrics were very similar, with a slight decrease ($\geq$ 0.02 AUC) in some settings. This preliminary step was essential to confirm the effectiveness of the used machine learning pipeline with sate-of-the-art results.

Predictive Performance

Fractal versus state-of-the-art features

Fractal features stood out by being included in the best-performing models for all the diseases for diagnosis, and the same for prognosis except for arrhythmia and valvular heart disease.

In the comparison of the best fractal and radiomics models, the ones for radiomics generally achieved slightly higher mean balanced accuracy, whereas the mean performance scores in the fractal models were slightly better for prognosis. Despite this trend, no statistical significant differences were observed in any disease and prediction type (prognosis or diagnosis). With the evidence of these results in the analyzed diseases, fractal and radiomics models are considered equivalent in terms of their predictive performance.

Prognosis versus Diagnosis Performance

Generally, prognosis performance results were lower compared to those for diagnosis. This is in line with our expectations, as the difficulty of prognosticating over diagnosing is well known. Diagnosis models predict the presence or absence of a disease based on the current evidence, whereas prognostic models have to predict future outcomes based on the current data. Predicting future events is a more complex task, especially because several factors that may arise in the future could influence the outcome and are not considered by the model, such as new lifestyle changes or medical treatments. This phenomenon can have a more negative impact on diseases that take a long time to develop, such as heart failure and valvular heart disease.

Impact of Including Additional Feature Sets

In the predictive performance results, it was observed that, on some occasions, the performance of a given model was not improved by adding additional features in its feature set. However, in cases where performance decreased, the decline was very small, with a maximum drop of 0.03 mean balanced accuracy. Moreover, in these situations, the standard deviations of the performance distribution in the testing cross-validation overlapped, and when these cases were further inspected with the Wilcoxon Signed-Rank non-parametric test, no significant differences were found. To thoroughly assess how using additional feature sets affects the performance, it might be beneficial to increase the number of testing folds, which could reduce the variance in performance and provide a more stable evaluation of the model. Also, the slight performance drop could be explained by the choice of the feature selection method, that might not have prioritized the most truly relevant features due to its reliance on model weights, as explained before.

Fairness Results

In the fairness results, focusing on the main fairness metric (i.e., the average odds difference), the best fractal and radiomics models could be considered similarly fair, as in most cases one model was better in two sensitive attributes while the other model had less bias in the remaining two attributes. In addition, the choice of the model in terms of fairness could differ depending on which sensitive attributes are considered more important. To make an informed decision, it is important to note that the best fractal models had less bias toward obesity and hypercholesterolemia, whereas radiomics models had less bias toward sex and age.

When applying mitigation with the exponentiated gradient mitigation technique, the fairness metrics of all sensitive attributes were substantially improved, which reflects the effectiveness of implementing that in-processing mitigation method. Also, the improvement in fairness was not accompanied by a large drop in performance, with a drop of less than 0.1 of balanced accuracy in most cases. Regarding the fractal and radiomics models comparison, fractals models were generally fairer after mitigation. Similarly, in the fairness results of the feature that combined the other sensitive attributes, radiomics models had better fairness values, while fractal suffer from less bias when applying fairness mitigation.

A plausible reason why mitigation had a more profound effect on models utilizing fractal features is their reduced number of features compared to the radiomics feature sets. This is because it might be easier for the exponentiated gradient method to find a feasible solution that balances predictive performance and fairness in its optimization when the model has fewer dimensions.

Feature Importance

In the feature importance results for diagnosis, apart from the clinical variables (age and sex), myocardium complexity and heterogeneity, as measured by its fractal dimension and lacunarity, respectively, were very important for almost all the considered cardiovascular diseases. In the prognosis scenario, excluding age and sex, myocardium complexity (assessed by the fractal dimension) was also very relevant. Moreover, the fractal features for the right and the left ventricle were present in the top 3 features for the same number of disease (2 in both cases). especially for stroke, where two of the top features belong to the right ventricle.

The finding that the right and left ventricle appear similarly in the significance of cardiovascular disease is surprising, as the traditional focus for clinical practice was to only check the left ventricle function, without paying attention to the right one. Because of this, the right ventricle is known as *the forgotten ventricle*. However, these results are in line with recent research that emphasizes the need for greater focus on the right ventricle(2022) [86].

6.1.2 Limitations and Assumptions

Censoring in the Prospective Follow-up

As introduced earlier, in the UK-Biobank project, participants incorporated in the cohort in a gradual manner and the CMR imaging visit date could be different for all of them. As there is only information up to a certain time limit (i.e., the last UK-Biobank update), the follow-up period differs between the first and the last participants who underwent CMR imaging, which was also evidenced by the high variability found in the follow-up time span, with more than 2 years of standard deviation. This variability could make that some subjects who were followed by a short period of time and did not suffer the target disease during their follow-up were used as a control case, they could have been diagnosed after the date of the last UK-Biobank update, even though the maximum follow-up time span would have captured that case as a positive one. Thus, the date of the last UK-Biobank update was used as a censoring date when it was more recent than the maximum follow-up date.

Limitations of the Machine Learning Pipeline

There are two main limitations of the machine learning pipeline that come from decisions that had to be made due to the long execution time of all experiments with the available computational resources, which would otherwise have meant several days for each disease and its corresponding diagnostic and prognostic scenarios. First, the number of folds used for testing was lower than the standard range (5 to 10 folds) commonly employed in research [87]. In addition, even though selecting the features based on model's weights is computational efficient, it could lead to not reaching the same performance as other feature selection methods that employ a more exhaustive approach and focus on direct performance improvement, such as sequential forward or backward feature selection.

Definition of the Control Group

In all the analyzed cardiovascular diseases, the control group was defined as participants who were just not diagnosed with the disease in the target time span (either the period up to the imaging visit for diagnosis, or the prospective follow-up). This means that the control group was not necessarily completely free from cardiovascular diseases, not having to represent actual healthy subjects. The rationale behind this definition of the control group was to mimic real-world clinical practice. Generally, medical imaging tests like CMR are only performed on people who suffer from other cardiovascular diseases or with suspicion of one of them. Thus, the interpretation of the results when including cardiovascular sick participants from other diseases could potentially help in the decision-making in the cardiologists' clinical practice. However, this is a much more complex scenario, as it is easier to distinguish between a completely healthy cardiovascular system compared to a diseased one. In addition, several diseases could share injuries and texture patterns of the ventricles or the myocardium, which makes it even more difficult to achieve a high predictive accuracy.

Limitations of the Statistical Testing

In the Wilcoxon Signed-Rank test, there were only 3 paired observations from the 3 folds of the outer cross-validation used for testing. With such a reduced number of performance samples, the power of the test to detect a true difference could be limited, since a sample size of just 3 values could not accurately reflect the true performance distribution of the models. For instance, even though the Wilcoxon Signed-Rank test is considered more robust to outliers compared to parametric tests by relying on ranks of the differences between pairs, with just 3 performance values, any outlier would have a noticeable impact on the ranking, potentially affecting the outcome of the statistical test.

Assumptions Related to Data Collection

Some assumptions were made in the data collection phase. For example, age was the chronological age in years at imaging visit, which represents the end of the time period to consider the disease diagnosis for the diagnosis scenario, and the start of the prospective follow-up. In the case of diagnosis, this means that the age could be later than the disease diagnosis, and the time difference between the patient's age and the diagnosis date could be noticeable if it occurred at the beginning of the time period, making it more difficult for models to predict accurately.

Moreover, although the variable indicating high cholesterol was not only used for fairness purposes and was not included in the main feature sets (fractals, radiomics, and CMR indices), the used criteria to determine high cholesterol could have some limitations. A positive case for high cholesterol was defined with participants taking cholesterol lowering drugs or having a serum cholesterol value greater than 7 mmol/L. First, being prescribed with cholesterol-lowering medications does necessarily involve having hypercholesterolemia, as patients could be prescribed these drugs as a preventative measure based on the family history or other risk factors. Regarding the cholesterol value threshold, a single blood test could not reflect the general cholesterol levels in a certain time period, as it represents a snapshot, and cholesterol levels could drastically change over time, both short-term (i.e., with food ingestion before testing) or long-term. In addition, there are different criteria to determine if a given level is considered low or high depending on the country or medical cardiology association. Also, it should be taken into account that the threshold to diagnose hypercholesterolemia from the same medical institution usually changes over time, as evidenced by the historical decrease in the thresholds. This suggests that the threshold used to determine a high cholesterol value may become outdated in the near future.

Generelizability of results

In the descriptive statistics, it was observed that the age of the study sample ranged from 44 to 82 years at the time of the imaging visit, with a median age of 64 years. Thus, infants, children, adolescents and young adults are not represented at all, and the results will be biased toward the cardiovascular disease patterns of middle and older adults, as well as seniors.

In the same line, an important limitation of the cohort is that most of the participants are which and are generally healthier for living in United Kingdom. Furthermore, people who voluntarily agree to participate in a research project like the UK Biobank tend to be healthier, which is known as the *healthy volunteer* effect. Thus, the number of cases was considerably low for some diseases, with the number of positive cases being lower than 200 for some of them, which could limit the model's ability to learn disease-associated patterns and have a negative impact on the predictive performance.

The results of this thesis should be generalized with caution taking into account all the assumptions and limitations mentioned above, as the diagnosis and prognosis of cardiovascular diseases involve human lives, which are of utmost importance.

6.2 Ethical-social impact, sustainability, and diversity

6.2.1 Ethical-social impact

A substantial part of this thesis has focused on assessing the fairness of the models by means of various sensitive attributes. These include gender, age and obesity. Considering the fairness of predictive models with these attributes draws attention to the biases of predictive models, which have remained hidden because they have traditionally been treated as black boxes. Going a step further, in this thesis we have not only evaluated the fairness metrics, but also mitigated the bias of the models, forcing them to be fairer. In this way, vulnerable groups such as women, elderly, and people suffering from obesity will suffer less discrimination when a prediction is made with their data, making the probability of a prognosis or diagnosis of a disease, as well as false positives or negatives, similar regardless of the sensitive attribute group to which the person belongs.

On the other hand, the use of machine learning models for the diagnosis and prognosis of cardiovascular diseases can help to improve the under-diagnosis of diseases and reduce medical costs in developing countries that unfortunately cannot devote so many resources to healthcare. Thus, machine learning algorithms could serve as democratisers of health, reducing ethical and social differences.

6.2.2 Sustainability

Using machine learning models could improve sustainability indirectly by helping to diagnose and prognose cardiovascular diseases. By improving the accuracy of early detection of diseases and reducing underdiagnosis, the associated energy expenditure that these conditions entail when they manifest suddenly (for example, through a heart attack) could be reduced, which could lead to long hospital stays.

On the other hand, the use of tabular features extracted from medical images, such as fractal features, could save computing time, since once extracted they can be reused, instead of dealing directly with heavy images like some deep learning models with medical high resolution images. This effect could even more pronounced with the type of fractal features proposed, since there is a very low number of features, which would mean even less expenditure of resources. This lower computational demand could reduce energy expenditure, positively contributing to a task as important as the health of our planet.

6.2.3 Diversity

One way to try to consider as much diversity as possible is to use a very large data set. In this thesis, a sample of more than 30,000 participants has been used, which probably includes very diverse people, with subjects of different genders and races. However, it must be taken into account that all participants come from United Kingdom, so, although there are a large number of groups represented, they could be represented in unequal proportions (for example, with more white people).

On the other hand, in research, a high-quality machine learning pipeline is often developed (with all the hard work that entails), but it is only used for one or a few diseases. On the other hand, this thesis has addressed the prediction of a large number of cardiovascular diseases, including both diagnosis and prognosis, with a total of 12 prediction scenarios. This approach advocates the diversity of diseases, being able to contribute positively to more (and more different) people, favoring diversity both in research and in people's lives.

6.3 Conclusions

In this thesis, the performance of machine learning models utilizing fractal-based features derived from CMR was evaluated for the diagnosis and prognosis of cardiovascular diseases considering both predictive performance and fairness. The target cardiovascular diseases were angina, arrhythmia, heart failure, myocardial infarction, stroke, and valvular heart disease. The study sample consisted of 31,931 subjects of the UK-Biobank who underwent CMR. The used fractal features included the fractal dimension and the lacunarity of each tissue (myocardium, left and right ventricle) and phase (end-systole, end-diastole).

Several machine learning models were utilized, including SVM, gradient boosting (XGboost, LightGBM, AdaBoost), and bagging (random forest, balanced random forest) models. The large imbalance between classes was addressed by random undersampling and the use of balanced accuracy as a performance metric. The machine learning pipeline involved a two-level 3-fold cross-validation (CV), with the inner CV for hyperparameter tuning (Bayesian-based) and the outer CV for testing. The performance of the fractal models was compared to widely-used features for cardiovascular diseases like conventional CMR indices and radiomics, and the Wilcoxon Signed-Rank test was conducted to determine statistical significant differences in the predictive performance resuts. Fairness was assessed with the demographic parity,

the equal opportunity difference, and the average odds difference (AOD), which was the main fairness metric. Fairness mitigation was applied through the exponentiated gradient mitigation technique. The similar performance obtained in the replication of state-of-the-art results confirmed the efficacy of the proposed machine learning pipeline. The best-performing fractal models were also inspected through the feature importance based on SHAP.

In the predictive performance results, no statistical significant differences were found between the mean balanced accuracy scores of the best fractal and radiomics models for each disease. Thus, it is considered that both models have an equivalent predictive power.

In addition, fractal features were in the feature set of the best overall performing models for all diseases in the diagnosis scenario and for all but arrhythmia and valvular heart disease in prognosis.

Regarding the fairness, the average odds difference showed that the fairness of the fractal and radiomics models was similar, with fractal models being less bias toward obesity and hypercholesterolemia, and radiomics models less bias toward sex and age. However, when mitigation was applied with the exponentiated gradient in-processing technique, fractal models were superior to radiomics models in almost every setting.

Furthermore, the number of fractal features is much smaller than that of radiomics, with the difference being in the order of hundreds. This large difference in feature set size could make models using radiomics more prone to overfitting as well as lead to much longer computation time or require feature selection methods that can limit performance. In addition, a low number of feature could be easier and faster to understand for a healthcare professional, that may be overwhelmed by knowing so many different radiomics features.

Therefore, fractals are proposed as an alternative to radiomics for future research, given their equivalent predictive performance in the used experiments, their lower bias if mitigation is applied, as well as their smaller number of features.

Regarding the SHAP feature importance results of the best fractal models, excluding clinical features like age and sex, the myocardium complexity and heterogeneity, as assessed by its fractal dimension and lacunarity, were very relevant for most of the diseases, especially in the diagnosis scenario. In addition, fractal features of the right ventricle was among the top three important features of two models for prognosis, which suggests that the right ventricle should not be as overlooked as it has historically been when prognosticating cardiovascular diseases.

Finally, it is important to consider that these results have been obtained from a UK-Biobank sample, which mainly consisted of white adult and senior participants and are healthier compared to other people who do not volunteer for studies or who live in countries with poor healthcare systems. Thus, the presented results should be extrapolated with caution to other populations.

Chapter 7

Future work

As future work, different statistical tests to compare the best fractal and radiomics models could be tried as an alternative to the Wilcoxon Signed-Rank test. If the same number of testing folds are used due to the computational resources limitations, one way to address the model comparison is by comparing their predictions, like McNemar test. McNemar test compares the predictions of two models by considering the cases in which they disagree. The predictions of all folds would be aggregated, and a single McNemar test would be conducted with all of them. As this test employs as many paired samples as the number of predicted observations (which are substantially more than the number of folds), it would likely have more power to detect significant differences if they really exist. However, the performance of the models would not be compared, but rather the degree of disagreement between the predictions of both models. Thus, the test could find statistical significant differences even if they actually achieve the same performance.

Moreover, it would be interesting to conduct the same experiments by using subjects completely free of cardiovascular conditions in the control group. Although it could be less directly applicable to the current clinical practice, the predictive models could achieve higher performance, because it is easier to distinguish a fully healthy participant from a sick one. Additionally, it would be very attractive from a research point of view, as the disease-specific signatures could be better captured. Also, the most important features of the models using both definitions of the control groups could be compared to see it they remain broadly unchanged.

Furthermore, the feature importance based on Shapley values could be used to elucidate the most important features of the best model combining the fractal and radiomics features sets. Examining the most relevant features in that model would represent a direct method to compare the impact of each feature type, by checking whether fractal-based features are positioned at the top, or instead all the most important features belong to the radiomics set. This would shed light to a more granular fractal and radiomics comparison, which would be at the feature level instead of generally comparing the feature sets. However, regarding the practical implementation of SHAP (the library for computing the feature importance based on Shapley values), it should be considered that it only supports kernel or permutation explainers for the AdaBoost model, which take a great deal of time (in the order of days).

Most of the identified limitations of this study are related to the restricted computational power available. A high-performance environment may be needed to conduct the required experiments, so that the choice of the methods used is not constrained by the computational capability at hand. With such an environment, the number of testing folds could be increased to five, and more exhaustive feature selection methods like Sequential Forward Feature Selection (SFFS) could be used, potentially leading to more stable predictions and higher performance.

In addition, although a wide variety of models have been utilized in this study (namely SVC, and several gradient boosting and bagging models), other models capable of capturing complex patterns have not been explored. Although tree-based models are generally considered superior for tabular data [88] [89], a field of research has emerged about the application of neural networks to this type of data, with some recent models like SAINT showing similar or even slightly better performance compared to tree-based models [90].

Also, the diagnosis and prognosis of different types of diseases beyond cardiovascular conditions could be investigated. The UK-Biobank provides the ICD-10 codes to represent patient diagnosis, including cancer-related diseases. Some researchers have addressed the potential of the application of medical imaging techniques like CMR to identify the cardiac toxicity due to cancer treatments [91] [92]. However, the relationships that some studies have found between cancer and heart disease not related to cancer treatment [93] have not been further explored. Some physiologically plausible explanations of this relationship are the creation of blood vessels that are created to support cancer growth (known as angiogenesis), and the early detection of metastasis in the myocardium.

Finally, future research could conduct similar experiments on the same diseases in cohorts with a larger number of positive cases, so that the sample size is not a limitation to the predictive power of the models.

Bibliography

[1] Gregory A Roth, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, and et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: A systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159):1736–1788, Nov 2018. https://doi.org/10.1016/s0140-6736(18)32203-7.

[2] Gregory A. Roth, George A. Mensah, Catherine O. Johnson, Giovanni Addolorato, Enrico Ammirati, Larry M. Baddour, Noël C. Barengo, Andrea Z. Beaton, Emelia J. Benjamin, Catherine P. Benziger, and et al. Global burden of cardiovascular diseases and risk factors, 1990–2019. *Journal of the American College of Cardiology*, 76(25):2982–3021, Dec 2020. https://doi.org/10.1016/j.jacc.2020.11.010.

[3] D. E. Bedford. The ancient art of feeling the pulse. *Heart*, 13(4):423–437, Oct 1951. https://doi.org/10.1136/hrt.13.4.423.

[4] Ibrahim R Hanna and Mark E Silverman. A history of cardiac auscultation and some of its contributors. *The American Journal of Cardiology*, 90(3):259–267, Aug 2002. https://doi.org/10.1016/s0002-9149(02)02465-7.

[5] Erik L Ritman. Cardiac computed tomography imaging: A history and some future possibilities. *Cardiology Clinics*, 21(4):491–513, Nov 2003. https://doi.org/10.1016/s0733-8651(03)00092-4.

[6] Tal Geva. Magnetic resonance imaging: Historical perspective. *Journal of Cardiovascular Magnetic Resonance*, 8(4):573–580, Aug 2006. https://doi.org/10.1080/10976640600755302.

[7] C. B. Marcu, A. M. Beek, and A. C. van Rossum. Clinical applications of cardiovascular magnetic resonance imaging. *Canadian Medical Association Journal*, 175(8):911–917, Oct 2006. https://doi.org/10.1503/cmaj.060566.

[8] Xiaoyue Zhou, Yucheng Chen, Rob J. van der Geest, Peng Hu, and Ming-Yen Ng. Editorial: Advanced quantitative indexes in cardiovascular magnetic resonance imaging. *Frontiers in Cardiovascular Medicine*, 11, Feb 2024. https://doi.org/10.3389/fcvm.2024.1302397.

[9] Nadine Kawel-Boehm, Scott J. Hetzel, Bharath Ambale-Venkatesh, Gabriella Captur, Christopher J. Francois, Michael Jerosch-Herold, Michael Salerno, Shawn D. Teague, Emanuela Valsangiacomo-Buechel, Rob J. van der Geest, and et al. Reference ranges ("normal values") for cardiovascular magnetic resonance (cmr) in adults and children: 2020 update. *Journal of Cardiovascular Magnetic Resonance*, 22(1):87, Jan 2020. https://doi.org/10.1186/s12968-020-00683-3.

[10] Marius E. Mayerhoefer, Andrzej Materka, Georg Langs, Ida Häggström, Piotr Szczypiński, Peter Gibbs, and Gary Cook. Introduction to radiomics. *Journal of Nuclear Medicine*, 61(4):488–495, Feb 2020. https://doi.org/10.2967/jnumed.118.222893.

[11] Zahra Raisi-Estabragh, Cristian Izquierdo, Victor M Campello, Carlos Martin-Isla, Akshay Jaggi, Nicholas C Harvey, Karim Lekadir, and Steffen E Petersen. Cardiac magnetic resonance radiomics: Basic principles and clinical perspectives. *European Heart Journal - Cardiovascular Imaging*, 21(4):349–356, Mar 2020. https://doi.org/10.1093/ehjci/jeaa028.

[12] Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. Radiomics: Images are more than pictures, they are data. *Radiology*, 278(2):563–577, Feb 2016. https://doi.org/10.1148/radiol.2015151169.

[13] R. W. Glenny, H. T. Robertson, S. Yamashiro, and J. B. Bassingthwaighte. Applications of fractal analysis to physiology. *Journal of Applied Physiology*, 70(6):2351–2367, Jun 1991. https://doi.org/10.1152/jappl.1991.70.6.2351.

[14] Gabriella Captur, Audrey L. Karperien, Chunming Li, Filip Zemrak, Catalina Tobon-Gomez, Xuexin Gao, David A. Bluemke, Perry M. Elliott, Steffen E. Petersen, and James C. Moon. Fractal frontiers in cardiovascular magnetic resonance: Towards clinical implementation. *Journal of Cardiovascular Magnetic Resonance*, 17(1):80, Jan 2015. https://doi.org/10.1186/s12968-015-0179-0.

[15] Elina T. Ziukelis, Elijah Mak, Maria-Eleni Dounavi, Li Su, and John T O'Brien. Fractal dimension of the brain in neurodegenerative disease and dementia: A systematic review. *Ageing Research Reviews*, 79:101651, Aug 2022. https://doi.org/10.1016/j.arr.2022.101651.

[16] Leticia Díaz Beltrán, Christopher R. Madan, Carsten Finke, Stephan Krohn, Antonio Di Ieva, and Francisco J. Esteban. Fractal dimension analysis in neurological disorders: An overview. *Advances in Neurobiology*, page 313–328, 2024. https://doi.org/10.1007/978-3-031-47606-8_16.

[17] Justin M. Zook and Khan M. Iftekharuddin. Statistical analysis of fractal-based brain tumor detection algorithms. *Magnetic Resonance Imaging*, 23(5):671–678, Jun 2005. https://doi.org/10.1016/j.mri.2005.04.002.

[18] Dheerendranath Battalapalli, Sreejith Vidyadharan, B. V. Prabhakar Rao, P. Yogeeswari, C. Kesavadas, and Venkateswaran Rajagopalan. Fractal dimension: Analyzing its potential as a neuroimaging biomarker for brain tumor diagnosis using machine learning. *Frontiers in Physiology*, 14, Jul 2023. https://doi.org/10.3389/fphys.2023.1201617.

[19] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, Apr 2021. https://doi.org/10.1007/s12525-021-00475-2.

[20] Michał Strzelecki and Pawel Badura. Machine learning for biomedical application. *Applied Sciences*, 12(4):2022, Feb 2022. https://doi.org/10.3390/app12042022.

[21] Hafsa Habehh and Suril Gohel. Machine learning in healthcare. *Current Genomics*, 22(4):291–300, Dec 2021. https://doi.org/10.2174/1389202922666210705124359.

[22] Shigao Huang, Jie Yang, Simon Fong, and Qi Zhao. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Letters*, 471:61–71, Feb 2020. https://doi.org/10.1016/j.canlet.2019.12.007.

[23] Min Yang, Qiqi Cao, Zhihan Xu, Yingqian Ge, Shujiao Li, Fuhua Yan, and Wenjie Yang. Development and validation of a machine learning-based radiomics model on cardiac computed tomography of epicardial adipose tissue in predicting characteristics and recurrence of atrial fibrillation. *Frontiers in Cardiovascular Medicine*, 9, Mar 2022. https://doi.org/10.3389/fcvm.2022.813085.

[24] Irem Cetin, Zahra Raisi-Estabragh, Steffen E. Petersen, Sandy Napel, Stefan K. Piechnik, Stefan Neubauer, Miguel A. Gonzalez Ballester, Oscar Camara, and Karim Lekadir. Radiomics signatures of cardiovascular risk factors in cardiac mri: Results from the uk biobank. *Frontiers in Cardiovascular Medicine*, 7, Nov 2020. https://doi.org/10.3389/fcvm.2020.591368.

[25] Esmeralda Ruiz Pujadas, Zahra Raisi-Estabragh, Liliana Szabo, Celeste McCracken, Cristian Izquierdo Morcillo, Víctor M. Campello, Carlos Martín-Isla, Angelica M. Atehortua, Hajnalka Vago, Bela Merkely, and et al. Prediction of incident cardiovascular events using machine learning and cmr radiomics. *European Radiology*, 33(5):3488–3500, Dec 2022. https://doi.org/10.1007/s00330-022-09323-z.

[26] Alvin Rajkomar, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169(12):866, Dec 2018. https://doi.org/10.7326/m18-1990.

[27] Vishwali Mhasawade, Yuan Zhao, and Rumi Chunara. Machine learning and algorithmic fairness in public and population health. *Nature Machine Intelligence*, 3(8):659–666, Jul 2021. https://doi.org/10.1038/s42256-021-00373-4.

[28] Gabriele A. Losa, Dušan Ristanović, Dejan Ristanović, Ivan Zaletel, and Stefano Beltraminelli. From fractal geometry to fractal analysis. *Applied Mathematics*, 7(4):346–354, Mar 2016. https://doi.org/10.4236/am.2016.74032.

[29] Rajendra Acharya U., P. Subbanna Bhat, N. Kannathal, Ashok Rao, and Choo Min Lim. Analysis of cardiac health using fractal dimension and wavelet transformation. *ITBM-RBM*, 26(2):133–139, Apr 2005. https://doi.org/10.1016/j.rbmret.2005.02.001.

[30] L A Neves, F R Oliveira, F A Peres, R D Moreira, A R Moriel, and F de. Maximum entropy, fractal dimension and lacunarity in quantification of cellular rejection in myocardial biopsy of patients submitted to heart transplantation. *Journal of physics. Conference series*, 285:012032–012032, Mar 2011. https://doi.org/10.1088/1742-6596/285/1/012032.

[31] Ateet Kosaraju, Amandeep Goyal, Yulia Grigorova, and Amgad N. Makaryus. Left ventricular ejection fraction. *StatPearls. NCBI Bookshelf.* https://www.ncbi.nlm.nih.gov/books/NBK459131/.

[32] Radiomic features. Pyradiomics documentation. https://pyradiomics.readthedocs.io/en/latest/features.html.

[33] William S Noble. What is a support vector machine? *Nature Biotechnology*, 24(12):1565–1567, Dec 2006. https://doi.org/10.1038/nbt1206-1565.

[34] Alessia Mammone, Marco Turchi, and Nello Cristianini. Support vector machines. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(3):283–289, Nov 2009. https://doi.org/10.1002/wics.49.

[35] Carl Kingsford and Steven L Salzberg. What are decision trees? *Nature Biotechnology*, 26(9):1011–1013, Sep 2008. https://doi.org/10.1038/nbt0908-1011.

[36] Barry de Ville. Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(6):448–455, Oct 2013. https://doi.org/10.1002/wics.1278.

[37] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), Feb 2018. https://doi.org/10.1002/widm.1249.

[38] Thomas G. Dietterich. Ensemble methods in machine learning. *Multiple Classifier Systems*, 1857:1–15, 2000. https://doi.org/10.1007/3-540-45014-9_1.

[39] Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. A survey on ensemble learning. *Frontiers of Computer Science*, 14(2):241–258, Aug 2019. https://doi.org/10.1007/s11704-019-8208-z.

[40] Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7(21), 2013. https://doi.org/10.3389/fnbot.2013.00021.

[41] Peter Bühlmann and Bin Yu. Boosting. *WIREs Computational Statistics*, 2(1):69–74, Dec 2009. https://doi.org/10.1002/wics.55.

[42] Xgboost 1.5.1 documentation. xgboost developers. https://xgboost.readthedocs.io/en/stable/.

[43] Lightgbm 3.3.5 documentation. lightgbm.readthedocs.io. https://lightgbm.readthedocs.io.

[44] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. https://doi.org/10.1006/jcss.1997.1504.

[45] Aakash Parmar, Rakesh Katariya, and Vatsal Patel. A review on random forest: An ensemble classifier. *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*, 26:758–763, Dec 2018. https://doi.org/10.1007/978-3-030-03146-6_86.

[46] A. S. More and Dipti P. Rana. Review of random forest classification techniques to resolve data imbalance. *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*, Oct 2017. https://doi.org/10.1109/icisim.2017.8122151.

[47] Balanced random forest classifier. *imbalanced-learn.org*. https://imbalanced-learn.org/stable/references/generated/imblearn.ensemble.BalancedRandomForestClassifier.html.

[48] Li Yang and Abdallah Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316, Nov 2020. https://doi.org/10.1016/j.neucom.2020.07.061.

[49] Tong Yu and Hong Zhu. Hyper-parameter optimization: A review of algorithms and applications. *arXiv.org*, Mar 2020. https://doi.org/10.48550/arXiv.2003.05689.

[50] Shuhei Watanabe. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *arXiv.org*, May 2023. https://doi.org/10.48550/arXiv.2304.11127.

[51] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, Jul 2019. https://doi.org/10.1145/3292500.3330701.

[52] Luca Oneto and Silvia Chiappa. Fairness in machine learning. *Recent Trends in Learning From Data*, page 155–196, 2020. https://doi.org/10.1007/978-3-030-43883-8_7.

[53] Richard J Chen, Judy J Wang, Drew, Tiffany Y Chen, Jana Lipkova, Ming Y Lu, Sharifa Sahai, and Faisal Mahmood. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering*, 7(6):719–742, Jun 2023. https://doi.org/10.1038/s41551-023-01056-8.

[54] Max Hort, Zhenpeng Chen, Jie M Zhang, Mark Harman, and Federica Sarro. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, Nov 2023. https://doi.org/10.1145/3631326.

[55] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, Aug 2023. https://doi.org/10.1145/3616865.

[56] Andreas Holzinger. From machine learning to explainable ai. *IEEE Xplore*, page 55–66, Aug 2018. https://doi.org/10.1109/DISA.2018.8490530.

[57] Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, 2020. https://doi.org/10.1109/access.2020.2976199.

[58] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *arXiv.org*, Nov 2017. https://doi.org/10.48550/arXiv.1705.07874.

[59] Paul Sajda. Machine learning for detection and diagnosis of disease. *Annual Review of Biomedical Engineering*, 8(1):537–565, Aug 2006. https://doi.org/10.1146/annurev.bioeng.8.061505.095802.

[60] Tim Leiner, Daniel Rueckert, Avan Suinesiaputra, Bettina Baeßler, Reza Nezafat, Ivana Išgum, and Alistair A. Young. Machine learning in cardiovascular magnetic resonance: Basic concepts and applications. *Journal of Cardiovascular Magnetic Resonance*, 21(1):61, Jan 2019. https://doi.org/10.1186/s12968-019-0575-y.

[61] Marleen de Bruijne. Machine learning approaches in medical image analysis: From detection to diagnosis. *Medical Image Analysis*, 33:94–97, Oct 2016. https://doi.org/10.1016/j.media.2016.06.032.

[62] Mahboobeh Jafari, Afshin Shoeibi, Marjane Khodatars, Navid Ghassemi, Parisa Moridian, Roohallah Alizadehsani, Abbas Khosravi, Sai Ho Ling, Niloufar Delfan, Yu-Dong Zhang, and et al. Automated diagnosis of cardiovascular diseases from cardiac magnetic resonance imaging using deep learning models: A review. *Computers in Biology and Medicine*, 160:106998, Jun 2023. https://doi.org/10.1016/j.compbiomed.2023.106998.

[63] Rhodri H. Davies, JoÃ£o B. Augusto, Anish Bhuva, Hui Xue, Thomas A. Treibel, Yang Ye, Rebecca K. Hughes, Wenjia Bai, Clement Lau, Hunain Shiwani, and et al. Precision measurement of cardiac structure and function in cardiovascular magnetic resonance using machine learning. *Journal of Cardiovascular Magnetic Resonance*, 24(1):16, Jan 2022. https://doi.org/10.1186/s12968-022-00846-4.

[64] Andrew J Swift, Haiping Lu, Johanna Uthoff, Pankaj Garg, Marcella Cogliano, Jonathan Taylor, Peter Metherall, Shuo Zhou, Christopher S Johns, Samer Alabed, and et al. A machine learning cardiac magnetic resonance approach to extract disease features and automate pulmonary arterial hypertension diagnosis. *European Heart Journal - Cardiovascular Imaging*, 22(2):236–245, Jan 2020. https://doi.org/10.1093/ehjci/jeaa001.

[65] Emine Sebnem Durmaz, Mert Karabacak, Burak Berksu Ozkara, Osman Aykan Kargın, Utku Raimoglu, Hasan Tokdil, Eser Durmaz, and Ibrahim Adaletli. Radiomics-based machine learning models in stemi: A promising tool for the prediction of major adverse cardiac events. *European Radiology*, 33(7):4611–4620, Jan 2023. https://doi.org/10.1007/s00330-023-09394-6.

[66] Elham Avard, Isaac Shiri, Ghasem Hajianfar, Hamid Abdollahi, Kiara Rezaei Kalantari, Golnaz Houshmand, Kianosh Kasani, Ahmad Bitarafan-rajabi, Mohammad Reza Deevband, Mehrdad Oveisi, and et al. Non-contrast cine cardiac magnetic resonance image radiomics features and machine learning algorithms for myocardial infarction detection. *Computers in Biology and Medicine*, 141:105145, Feb 2022. https://doi.org/10.1016/j.compbiomed.2021.105145.

[67] Jie Wang, Yuancheng Li, Fuyao Yang, Laura Bravo, Ke Wan, Yuanwei Xu, Wei Cheng, Jiayu Sun, Yanjie Zhu, Tingxi Zhu, and et al. Fractal analysis: Prognostic value of left ventricular trabecular complexity cardiovascular mri in participants with hypertrophic cardiomyopathy. *Radiology*, 298(1):71–79, Jan 2021. https://doi.org/10.1148/radiol.2020202261.

[68] Wen-Yi Jiang, Bing-Hua Chen, Chen Zhang, Ruo-Yang Shi, Rui Wu, Dong-Aolei An, Xiao-Hai Ma, Luke Wesemann, Jiani Hu, Yan Zhou, and et al. Fractal analysis in cardiovascular magnetic resonance: Prognostic value of biventricular trabecular complexity in hypertrophic cardiomyopathy. *Cardiovascular Diagnosis and Therapy*, 13(6):1030–1042, Dec 2023. https://doi.org/10.21037/cdt-23-162.

[69] Marjan Firouznia, Albert K. Feeny, Michael A. LaBarbera, Meghan McHale, Catherine Cantlay, Natalie Kalfas, Paul Schoenhagen, Walid Saliba, Patrick Tchou, John Barnard, and et al. Machine learning–derived fractal features of shape and texture of the left atrium and pulmonary veins from cardiac computed tomography scans are associated with risk of recurrence of atrial fibrillation postablation. *Circulation: Arrhythmia and Electrophysiology*, 14(3), Mar 2021. https://doi.org/10.1161/circep.120.009265.

[70] Tanja Kurzendorfer, Katharina Breininger, Stefan Steidl, Alexander Brost, Christoph Forman, and Andreas Maier. Myocardial scar segmentation in lge-mri using fractal analysis and random forest classification. *2018 24th International Conference on Pattern Recognition (ICPR)*, Aug 2018. https://doi.org/10.1109/icpr.2018.8545636.

[71] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, and et al. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3), Mar 2015. https://doi.org/10.1371/journal.pmed.1001779.

[72] Chinmaya Panigrahy, Ayan Seal, Nihar Kumar Mahato, and Debotosh Bhattacharjee. Differential box counting methods for estimating fractal dimension of gray-scale images:

A survey. *Chaos, Solitons Fractals*, 126:178–202, Sep 2019. https://doi.org/10.1016/j.chaos.2019.06.007.

[73] Pinliang Dong. Test of a new lacunarity estimation method for image texture analysis. 21(17):3369–3373, Jan 2000. https://doi.org/10.1080/014311600750019985.

[74] Elisa Rauseo, Cristian Izquierdo Morcillo, Zahra Raisi-Estabragh, Polyxeni Gkontra, Nay Aung, Karim Lekadir, and Steffen E. Petersen. New imaging signatures of cardiac alterations in ischaemic heart disease and cerebrovascular disease using cmr radiomics. *Frontiers in Cardiovascular Medicine*, 8, Sep 2021. https://doi.org/10.3389/fcvm.2021.716577.

[75] Taeho Jo. Data encoding. *Springer eBooks*, page 47–68, Dec 2020. https://doi.org/10.1007/978-3-030-65900-4_3.

[76] Vinod Sharma. A study on data scaling methods for machine learning. *International Journal for Global Academic Scientific Research*, 1(1), Feb 2022. https://doi.org/10.55938/ijgasr.v1i1.4.

[77] Andrea Dal Pozzolo, Olivier Caelen, and Gianluca Bontempi. When is undersampling effective in unbalanced classification tasks? *Machine Learning and Knowledge Discovery in Databases*, page 200–215, 2015. https://doi.org/10.1007/978-3-319-23528-8_13.

[78] Selectfrommodel. Scikit-learn 0.23.1 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFromModel.html.

[79] Avoiding overfitting of decision trees. *Principles of Data Mining*, page 119–134. https://doi.org/10.1007/978-1-84628-766-4_8.

[80] Adaboost classifier. Scikit-learn 0.23.1 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html.

[81] Amalia Luque, Alejandro Carrasco, Alejandro Martín, and Ana de las Heras. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91:216–231, Jul 2019. https://doi.org/10.1016/j.patcog.2019.02.023.

[82] R. F. Woolson. Wilcoxon signed-rank test. *Encyclopedia of Biostatistics*, Jul 2005. https://doi.org/10.1002/0470011815.b2a15177.

[83] Tiago P. Pagano, Rafael B. Loureiro, Fernanda V. N. Lisboa, Rodrigo M. Peixoto, Guilherme A. S. Guimarães, Gustavo O. R. Cruz, Maira M. Araujo, Lucas L. Santos, Marco A. S. Cruz, Ewerton L. S. Oliveira, Ingrid Winkler, and Erick G. S. Nascimento. Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, 7(1):15, Mar 2023. https://doi.org/10.3390/bdcc7010015.

[84] Pratyush Garg, John Villasenor, and Virginia Foggo. Fairness metrics: A comparative analysis. *IEEE Xplore*, page 3662–3666, Dec 2020. https://doi.org/10.1109/BigData50022.2020.9378025.

[85] Alfa Yohannis and Dimitrios S Kolovos. Towards model-based bias mitigation in machine learning. Oct 2022. https://doi.org/10.1145/3550355.3552401.

[86] Myriam Amsallem, Olaf Mercier, Yukari Kobayashi, Kegan Moneghetti, and Francois Haddad. Forgotten no more. *JACC: Heart Failure*, 6(11):891–903, Nov 2018. https://doi.org/10.1016/j.jchf.2018.05.022.

[87] Payam Refaeilzadeh, Lei Tang, and Huan Liu. Cross-validation. *Encyclopedia of Database Systems*, page 1–7, 2016. https://doi.org/10.1007/978-1-4899-7993-3_565-2.

[88] Leo Grinsztajn, Edouard Oyallon, and Gael Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? 35:507–520, 2022. https://proceedings.neurips.cc/paper_files/paper/2022/file/0378c7692da36807bdec87ab043cdadc-Paper-Datasets_and_Benchmarks.pdf.

[89] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, May 2022. https://doi.org/10.1016/j.inffus.2021.11.011.

[90] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7499–7519, Jun 2024. https://doi.org/10.1109/tnnls.2022.3229161.

[91] Aaron Soufer and Lauren A. Baldassarre. The role of cardiac magnetic resonance imaging to detect cardiac toxicity from cancer therapeutics. *Current Treatment Options in Cardiovascular Medicine*, 21(6), May 2019. https://doi.org/10.1007/s11936-019-0732-5.

[92] Paaladinesh Thavendiranathan, Bernd J. Wintersperger, Scott D. Flamm, and Thomas H. Marwick. Cardiac mri in the assessment of cardiac injury and toxicity from cancer chemotherapy. *Circulation: Cardiovascular Imaging*, 6(6):1080–1091, Nov 2013. https://doi.org/10.1161/circimaging.113.000899.

[93] Rudolf A. de Boer, Wouter C. Meijers, Peter van der Meer, and Dirk J. van Veldhuisen. Cancer and heart disease: Associations and relations. *European Journal of Heart Failure*, 21(12):1515–1525, Jul 2019. https://doi.org/10.1002/ejhf.1539.

Annexes

Annex A. Correspondence between ICD-10 codes and component outcomes

Component outcome	ICD-10 codes
Arrhythmia	I48, I49, I49.0, I49.1, I49.2, I49.3, I49.4, I49.5, I49.8, I49.9
Angina	I20, I20.0, I20.1, I20.8, I20.9
Myocardial infarction	I21, I21.0, I21.1, I21.2, I21.3, I21.4, I21.9
Heart failure	I50, I50.0, I50.1, I50.9
Stroke	I60, I60.0, I60.1, I60.2, I60.3, I60.4, I60.5, I60.6, I60.7, I60.8, I60.9, I61, I61.0, I61.1, I61.2, I61.3, I61.4, I61.5, I61.6, I61.8, I61.9, I62, I62.0, I62.1, I62.9, I63, I63.0, I63.1, I63.2, I63.3, I63.4, I63.5, I63.6, I63.8, I63.9, I64
Valvular Heart Disease	I05, I05.0, I05.1, I05.2, I05.8, I05.9, I06, I06.0, I06.1, I06.2, I06.8, I06.9, I07, I07.0, I07.1, I07.2, I07.8, I07.9, I08, I08.0, I08.1, I08.2, I08.3, I08.8, I08.9, I34, I34.0, I34.1, I34.2, I34.8, I34.9, I35, I35.0, I35.1, I35.2, I35.8, I35.9, I36, I36.0, I36.1, I36.2, I36.8, I36.9, I37, I37.0, I37.1, I37.2, I37.8, I37.9

Correspondence between ICD-10 codes and component outcomes.

Annex B. Hyperparameter ranges of all the models.

SVM

Hyperparameter name	Range / Categories
C	[0.01, 10]
gamma	[0.001, 1.0]
kernel	rbf, poly, sigmoid

Hyperparameter ranges of the SVM classifier.

XGBoost

Hyperparameter name	Range / Categories
n_estimators	[10, 100]
max_depth	[3,10]
learning_rate	[0.01, 0.2]
min_child_weight	[1, 10]
gamma	[0.1, 5]
subsample	[0.6, 1.0]
colsample_bytree	[0.6, 1.0]

Hyperparameter ranges of the XGBoost classifier.

LightGBM

Hyperparameter name	Range / Categories
n_estimators	[10, 100]
max_depth	[3,10]
learning_rate	[0.01, 0.2]
num_leaves	[5, 30]
min_data_in_leaf	[20, 50]
min_gain_to_split	[0.001, 1.0]

Hyperparameter ranges of the LightGBM classifier.

AdaBoost

Hyperparameter name	Range / Categories
n_estimators	[10, 100]
max_depth	[3,10]
learning_rate	[0.01, 2.0]
algorithm	SAMME, SAMME.R

Hyperparameter ranges of the AdaBoost classifier.

Random Forest and Balanced Random Forest

Hyperparameter name	Range / Categories
n_estimators	[10, 100]
max_depth	[3,10]
max_features	sqrt, log2
min_samples_split	[2, 10]

Hyperparameter ranges of the Random Forest and Balanced Random Forest classifiers.

Annex C. Predictive performance results for every component outcome.

Diagnosis

Heart failure

	CMR	Rad	Frac	CMR+Rad	CMR+Frac	Rad+Frac	All
XGB	0.713 ±0.01	0.699 ±0.06	0.688 ±0.06	0.723 ±0.03	0.704 ±0.06	0.696 ±0.05	**0.745 ± 0.01**
Ada	0.702 ±0.05	0.693 ±0.03	0.672 ±0.02	0.696 ±0.02	0.710 ±0.04	0.711 ±0.04	**0.712 ± 0.02**
LGBM	0.699 ±0.05	0.702 ±0.05	0.680 ±0.02	0.730 ±0.04	0.723 ±0.05	0.691 ±0.03	**0.748 ± 0.04**
RF	0.709 ±0.01	**0.739 ± 0.05**	0.702 ±0.02	0.719 ±0.03	0.717 ±0.03	0.736 ±0.07	0.719 ±0.02
BRF	0.698 ±0.01	0.732 ±0.06	0.707 ±0.02	0.734 ±0.04	0.721 ±0.02	**0.737 ± 0.06**	0.730 ±0.06
SVM	0.685 ±0.04	0.688 ±0.04	0.657 ±0.02	0.690 ±0.03	0.692 ±0.04	0.691 ±0.05	**0.707 ± 0.04**

Predictive performance results for heart failure diagnosis.

Myocardial infarction

	CMR	Rad	Frac	CMR+Rad	CMR+Frac	Rad+Frac	All
XGB	0.690 ±0.02	0.698 ±0.01	0.704 ±0.01	**0.714 ± 0.02**	0.691 ±0.04	0.698 ±0.00	0.706 ±0.02
Ada	0.705 ±0.01	**0.707 ± 0.02**	0.698 ±0.01	0.699 ±0.01	0.705 ±0.02	0.698 ±0.01	0.700 ±0.03
LGBM	0.704 ±0.02	0.694 ±0.02	0.695 ±0.02	0.714 ±0.02	0.705 ±0.03	0.699 ±0.01	**0.715 ± 0.01**
RF	**0.704 ± 0.02**	0.677 ±0.02	0.681 ±0.02	0.679 ±0.02	0.694 ±0.02	0.675 ±0.01	0.676 ±0.02
BRF	**0.708 ± 0.01**	0.694 ±0.01	0.696 ±0.02	0.695 ±0.03	0.699 ±0.03	0.688 ±0.02	0.688 ±0.01
SVM	0.673 ±0.02	0.669 ±0.02	0.670 ±0.02	0.673 ±0.02	**0.674 ± 0.03**	0.671 ±0.00	0.669 ±0.02

Performance of all models and feature combinations for myocardial infarction diagnosis.

Stroke

	CMR	Rad	Frac	CMR+Rad	CMR+Frac	Rad+Frac	All
XGB	0.570 ±0.03	0.589 ±0.04	0.589 ±0.00	0.584 ±0.04	**0.600 ± 0.05**	0.581 ±0.04	0.577 ±0.04
Ada	**0.595 ± 0.07**	0.570 ±0.04	0.577 ±0.03	0.578 ±0.04	0.566 ±0.01	0.582 ±0.04	0.566 ±0.02
LGBM	0.579 ±0.04	0.587 ±0.03	0.578 ±0.01	0.590 ±0.03	0.572 ±0.02	0.592 ±0.02	**0.598 ± 0.03**
RF	0.568 ±0.02	0.597 ±0.04	0.586 ±0.01	0.568 ±0.05	0.587 ±0.05	**0.606 ± 0.04**	0.581 ±0.03
BRF	0.597 ±0.07	0.580 ±0.04	0.567 ±0.03	0.584 ±0.04	0.568 ±0.03	**0.608 ± 0.02**	0.586 ±0.03
SVM	0.556 ±0.06	0.562 ±0.04	0.549 ±0.02	0.558 ±0.04	0.556 ±0.02	**0.575 ± 0.03**	0.560 ±0.03

Performance of all models and feature combinations for stroke diagnosis.

Valvular heart disease

	CMR	Rad	Frac	CMR+Rad	CMR+Frac	Rad+Frac	All
XGB	0.650 ±0.01	0.640 ±0.03	**0.652 ± 0.02**	0.637 ±0.04	0.630 ±0.04	0.646 ±0.03	0.650 ±0.03
Ada	0.617 ±0.03	**0.647 ± 0.01**	0.624 ±0.02	0.626 ±0.02	0.595 ±0.03	0.619 ±0.01	0.644 ±0.03
LGBM	0.620 ±0.01	**0.655 ± 0.02**	0.621 ±0.03	0.638 ±0.02	0.634 ±0.01	0.653 ±0.01	0.627 ±0.02
RF	0.638 ±0.04	**0.656 ± 0.03**	0.623 ±0.04	0.635 ±0.00	0.633 ±0.02	0.645 ±0.02	0.640 ±0.03
BRF	0.651 ±0.02	0.640 ±0.01	0.638 ±0.03	0.640 ±0.03	**0.672 ± 0.01**	0.649 ±0.01	0.661 ±0.03
SVM	0.600 ±0.03	**0.621 ± 0.02**	0.602 ±0.02	0.617 ±0.01	0.601 ±0.03	0.613 ±0.02	0.612 ±0.03

Performance of all models and feature combinations for valvular heart disease diagnosis.

Prognosis

Heart failure

	CMR	Rad	Frac	CMR+Rad	CMR+Frac	Rad+Frac	All
XGB	**0.731 ± 0.02**	0.722 ±0.02	0.677 ±0.01	0.703 ±0.02	0.717 ±0.01	0.718 ±0.02	0.703 ±0.01
Ada	0.696 ±0.03	0.725 ±0.02	0.677 ±0.02	0.698 ±0.03	0.703 ±0.02	**0.727 ± 0.02**	0.711 ±0.02
LGBM	0.722 ±0.01	0.719 ±0.02	0.677 ±0.02	0.719 ±0.02	**0.724 ± 0.01**	0.717 ±0.01	0.719 ±0.01
RF	**0.721 ± 0.01**	0.706 ±0.02	0.683 ±0.01	0.693 ±0.02	0.719 ±0.01	0.696 ±0.01	0.692 ±0.01
BRF	0.719 ±0.01	0.712 ±0.01	0.690 ±0.02	0.716 ±0.03	**0.732 ± 0.01**	0.708 ±0.01	0.720 ±0.04
SVM	0.684 ±0.01	**0.696 ± 0.02**	0.662 ±0.01	0.686 ±0.02	0.694 ±0.02	0.686 ±0.02	0.684 ±0.03

Performance of all models and feature combinations for heart failure prognosis.

Myocardial infarction

	CMR	Rad	Frac	CMR+Rad	CMR+Frac	Rad+Frac	All
XGB	**0.641 ± 0.02**	0.612 ±0.02	0.637 ±0.01	0.615 ±0.02	0.632 ±0.02	0.625 ±0.03	0.612 ±0.03
Ada	0.622 ±0.04	0.595 ±0.01	**0.640 ± 0.01**	0.594 ±0.05	0.628 ±0.02	0.611 ±0.01	0.604 ±0.03
LGBM	0.613 ±0.03	0.604 ±0.02	**0.632 ± 0.01**	0.610 ±0.01	0.627 ±0.02	0.617 ±0.03	0.598 ±0.02
RF	0.623 ±0.01	0.612 ±0.02	0.619 ±0.02	0.596 ±0.03	**0.636 ± 0.02**	0.604 ±0.03	0.614 ±0.03
BRF	0.650 ±0.01	0.608 ±0.01	0.628 ±0.03	0.628 ±0.02	**0.661 ± 0.02**	0.629 ±0.01	0.632 ±0.03
SVM	0.595 ±0.03	0.587 ±0.02	0.602 ±0.01	0.588 ±0.02	**0.603 ± 0.02**	0.599 ±0.02	0.592 ±0.02

Performance of all models and feature combinations for myocardial infarction prognosis.

Stroke

	CMR	Rad	Frac	CMR+Rad	CMR+Frac	Rad+Frac	All
XGB	0.620 ±0.04	0.595 ±0.02	0.601 ±0.01	0.615 ±0.01	**0.629 ± 0.03**	0.610 ±0.03	0.609 ±0.03
Ada	0.610 ±0.05	0.596 ±0.03	**0.619 ± 0.04**	0.590 ±0.02	0.595 ±0.03	0.591 ±0.04	0.609 ±0.00
LGBM	0.632 ±0.03	0.589 ±0.04	0.600 ±0.03	0.620 ±0.02	**0.633 ± 0.03**	0.595 ±0.03	0.607 ±0.01
RF	0.636 ±0.02	0.581 ±0.03	0.615 ±0.02	0.612 ±0.02	**0.641 ± 0.00**	0.586 ±0.03	0.615 ±0.02
BRF	**0.638 ± 0.01**	0.602 ±0.03	0.576 ±0.02	0.611 ±0.04	0.637 ±0.01	0.581 ±0.01	0.612 ±0.01
SVM	0.585 ±0.03	0.575 ±0.03	0.564 ±0.03	0.580 ±0.01	**0.587 ± 0.03**	0.574 ±0.03	0.585 ±0.02

Performance of all models and feature combinations for stroke prognosis.

Valvular heart disease

	CMR	Rad	Frac	CMR+Rad	CMR+Frac	Rad+Frac	All
XGB	**0.688 ± 0.03**	0.674 ±0.01	0.685 ±0.02	**0.688 ± 0.01**	0.685 ±0.04	0.679 ±0.03	0.684 ±0.02
Ada	0.691 ±0.04	0.676 ±0.02	0.679 ±0.03	0.685 ±0.02	0.688 ±0.04	0.673 ±0.02	**0.695 ± 0.03**
LGBM	0.685 ±0.03	0.681 ±0.02	0.662 ±0.02	**0.689 ± 0.02**	**0.689 ± 0.03**	0.671 ±0.02	0.686 ±0.01
RF	**0.698 ± 0.03**	0.665 ±0.03	0.676 ±0.01	0.685 ±0.01	0.690 ±0.03	0.659 ±0.03	0.671 ±0.01
BRF	**0.690 ± 0.01**	0.667 ±0.01	0.683 ±0.02	0.665 ±0.01	**0.690 ± 0.02**	0.663 ±0.00	0.664 ±0.02
SVM	**0.668 ± 0.03**	0.657 ±0.01	0.650 ±0.01	0.659 ±0.02	0.662 ±0.03	0.653 ±0.01	0.657 ±0.02

Performance of all models and feature combinations for valvular heart disease prognosis.

Annex D. Fairness results for every component outcome.

Diagnosis

Arrhythmia

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	-0.510 ± (0.04)	**0.180 ± (0.04)**	**0.150 ± (0.04)**	0.387 ± (0.08)
Eq. Opp. Diff. (Rad)	**-0.367 ± (0.08)**	0.185 ± (0.02)	0.188 ± (0.01)	**0.347 ± (0.12)**
Avg. Odds. Diff. (Frac)	-0.465 ± (0.043)	**0.151 ± (0.04)**	**0.118 ± (0.03)**	0.356 ± (0.06)
Avg. Odds. Diff. (Rad)	**-0.345 ± (0.07)**	0.169 ± (0.02)	0.166 ± (0.01)	**0.284 ± (0.08)**
Dem. Parity Diff. (Frac)	-0.426 ± (0.05)	0.171 ± (0.02)	**0.124 ± (0.00)**	0.331 ± (0.03)
Dem. Parity Diff. (Rad)	**-0.335 ± (0.05)**	**0.164 ± (0.02)**	0.148 ± (0.02)	**0.230 ± (0.05)**

Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) for arrhythmia diagnosis.

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	**-0.076 ± (0.03)**	**0.031 ± (0.01)**	**0.034 ± (0.03)**	**0.055 ± (0.03)**
Eq. Opp. Diff. (Rad)	-0.176 ± (0.04)	0.132 ± (0.02)	0.124 ± (0.02)	0.161 ± (0.09)
Avg. Odds. Diff. (Frac)	**-0.029 ± (0.04)**	**0.003 ± (0.02)**	**0.012 ± (0.02)**	**0.034 ± (0.02)**
Avg. Odds. Diff. (Rad)	-0.154 ± (0.03)	0.118 ± (0.03)	0.102 ± (0.02)	0.111 ± (0.06)
Dem. Parity Diff. (Frac)	**-0.024 ± (0.00)**	**0.013 ± (0.01)**	**0.007 ± (0.01)**	**0.025 ± (0.01)**
Dem. Parity Diff. (Rad)	-0.149 ± (0.03)	0.134 ± (0.03)	0.110 ± (0.01)	0.069 ± (0.03)
Bal. Acc. Drop (Frac)	0.048	0.111	0.112	0.046
Bal. Acc. Drop (Rad)	**0.002**	**0.014**	**0.003**	**0.023**

Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) when applying mitigation (through exponentiated gradient) for arrhythmia diagnosis.

Heart failure

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	**-0.407 ± (0.06)**	0.168 ± (0.03)	0.169 ± (0.05)	0.344 ± (0.10)
Eq. Opp. Diff. (Rad)	-0.409 ± (0.09)	**0.141 ± (0.01)**	**0.157 ± (0.00)**	**0.131 ± (0.09)**
Avg. Odds. Diff. (Frac)	-0.374 ± (0.05)	0.116 ± (0.05)	0.124 ± (0.07)	0.277 ± (0.09)
Avg. Odds. Diff. (Rad)	**-0.328 ± (0.07)**	**0.093 ± (0.04)**	**0.095 ± (0.02)**	**0.071 ± (0.07)**
Dem. Parity Diff. (Frac)	-0.363 ± (0.02)	0.160 ± (0.02)	**0.132 ± (0.02)**	0.228 ± (0.08)
Dem. Parity Diff. (Rad)	**-0.337 ± (0.03)**	**0.122 ± (0.02)**	0.158 ± (0.00)	**0.045 ± (0.02)**

Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) for heart failure diagnosis.

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	**-0.135 ± (0.07)**	0.153 ± (0.17)	**0.049 ± (0.01)**	0.127 ± (0.02)
Eq. Opp. Diff. (Rad)	-0.402 ± (0.09)	**0.139 ± (0.01)**	0.151 ± (0.00)	**0.120 ± (0.09)**
Avg. Odds. Diff. (Frac)	**-0.009 ± (0.07)**	**0.060 ± (0.1)**	**0.002 ± (0.02)**	0.070 ± (0.02)
Avg. Odds. Diff. (Rad)	-0.312 ± (0.07)	0.091 ± (0.04)	0.073 ± (0.02)	**0.057 ± (0.07)**
Dem. Parity Diff. (Frac)	**-0.019 ± (0.01)**	**0.006 ± (0.00)**	**0.004 ± (0.00)**	**0.021 ± (0.01)**
Dem. Parity Diff. (Rad)	-0.329 ± (0.03)	0.118 ± (0.02)	0.129 ± (0.00)	0.034 ± (0.02)
Bal. Acc. Drop (Frac)	0.135	0.131	0.183	0.063
Bal. Acc. Drop (Rad)	**0.052**	**0.052**	**0.081**	**0.024**

Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) when applying mitigation (through exponentiated gradient) for heart failure diagnosis.

Myocardial infarction

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	-0.775 ± (0.1)	**0.164 ± (0.03)**	**0.138 ± (0.01)**	0.289 ± (0.07)
Eq. Opp. Diff. (Rad)	**-0.656 ± (0.06)**	0.167 ± (0.04)	0.227 ± (0.03)	**0.207 ± (0.09)**
Avg. Odds. Diff. (Frac)	-0.705 ± (0.08)	**0.096 ± (0.04)**	**0.089 ± (0.04)**	0.267 ± (0.07)
Avg. Odds. Diff. (Rad)	**-0.604 ± (0.09)**	0.104 ± (0.05)	0.196 ± (0.04)	**0.180 ± (0.07)**
Dem. Parity Diff. (Frac)	-0.639 ± (0.07)	**0.170 ± (0.03)**	**0.140 ± (0.01)**	0.282 ± (0.08)
Dem. Parity Diff. (Rad)	**-0.556 ± (0.11)**	0.175 ± (0.04)	0.210 ± (0.01)	**0.173 ± (0.04)**

Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) for myocardial infarction diagnosis.

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	**-0.275 ± (0.18)**	0.154 ± (0.03)	**0.119 ± (0.01)**	0.168 ± (0.07)
Eq. Opp. Diff. (Rad)	-0.478 ± (0.19)	**0.151 ± (0.02)**	0.182 ± (0.04)	**0.141 ± (0.00)**
Avg. Odds. Diff. (Frac)	**-0.224 ± (0.21)**	**0.076 ± (0.05)**	**0.085 ± (0.03)**	**0.117 ± (0.10)**
Avg. Odds. Diff. (Rad)	-0.417 ± (0.18)	0.094 ± (0.04)	0.157 ± (0.04)	0.124 ± (0.01)
Dem. Parity Diff. (Frac)	**-0.249 ± (0.16)**	**0.159 ± (0.03)**	**0.121 ± (0.00)**	0.136 ± (0.09)
Dem. Parity Diff. (Rad)	-0.362 ± (0.17)	**0.159 ± (0.02)**	0.178 ± (0.03)	**0.125 ± (0.03)**
Bal. Acc. Drop (Frac)	0.061	**0.003**	**0.001**	0.013
Bal. Acc. Drop (Rad)	**0.02**	**0.003**	0.013	**0.012**

Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) when applying mitigation (through exponentiated gradient) for myocardial infarction diagnosis.

Stroke

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	-0.674 ± (0.16)	0.226 ± (0.03)	**0.207 ± (0.06)**	0.537 ± (0.15)
Eq. Opp. Diff. (Rad)	**-0.386 ± (0.02)**	**0.195 ± (0.03)**	0.268 ± (0.08)	**0.213 ± (0.09)**
Avg. Odds. Diff. (Frac)	-0.580 ± (0.2)	0.175 ± (0.04)	**0.063 ± (0.10)**	0.500 ± (0.15)
Avg. Odds. Diff. (Rad)	**-0.301 ± (0.04)**	**0.145 ± (0.06)**	0.224 ± (0.06)	**0.131 ± (0.05)**
Dem. Parity Diff. (Frac)	-0.489 ± (0.25)	0.222 ± (0.03)	**0.106 ± (0.02)**	0.491 ± (0.18)
Dem. Parity Diff. (Rad)	**-0.386 ± (0.02)**	**0.186 ± (0.04)**	0.230 ± (0.05)	**0.096 ± (0.06)**

Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) for stroke diagnosis.

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	**-0.189 ± (0.14)**	0.202 ± (0.05)	**0.185 ± (0.1)**	**0.170 ± (0.08)**
Eq. Opp. Diff. (Rad)	-0.277 ± (0.10)	**0.187 ± (0.03)**	0.243 ± (0.08)	0.191 ± (0.10)
Avg. Odds. Diff. (Frac)	**-0.126 ± (0.11)**	0.166 ± (0.08)	**0.055 ± (0.11)**	**0.096 ± (0.05)**
Avg. Odds. Diff. (Rad)	-0.210 ± (0.06)	**0.130 ± (0.06)**	0.195 ± (0.07)	0.115 ± (0.06)
Dem. Parity Diff. (Frac)	**-0.088 ± (0.06)**	0.191 ± (0.04)	**0.091 ± (0.02)**	0.160 ± (0.09)
Dem. Parity Diff. (Rad)	-0.277 ± (0.10)	**0.180 ± (0.04)**	0.205 ± (0.03)	**0.090 ± (0.04)**
Bal. Acc. Drop (Frac)	0.024	0.018	0.013	0.032
Bal. Acc. Drop (Rad)	**0.014**	**0.01**	**0.006**	**0.002**

Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) when applying mitigation (through exponentiated gradient) for stroke diagnosis.

Valvular heart disease

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	**-0.365 ± (0.07)**	**0.226 ± (0.00)**	**0.224 ± (0.09)**	0.436 ± (0.07)
Eq. Opp. Diff. (Rad)	-0.483 ± (0.08)	0.265 ± (0.03)	0.255 ± (0.09)	**0.116 ± (0.04)**
Avg. Odds. Diff. (Frac)	**-0.322 ± (0.06)**	0.210 ± (0.01)	**0.146 ± (0.06)**	0.408 ± (0.07)
Avg. Odds. Diff. (Rad)	-0.423 ± (0.06)	**0.200 ± (0.02)**	0.200 ± (0.07)	**0.090 ± (0.03)**
Dem. Parity Diff. (Frac)	**-0.280 ± (0.05)**	0.198 ± (0.02)	**0.105 ± (0.03)**	0.382 ± (0.07)
Dem. Parity Diff. (Rad)	-0.390 ± (0.01)	**0.141 ± (0.02)**	0.191 ± (0.01)	**0.066 ± (0.03)**

Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) for valvular heart disease diagnosis.

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	**-0.276 ± (0.05)**	0.219 ± (0.04)	**0.169 ± (0.1)**	0.254 ± (0.02)
Eq. Opp. Diff. (Rad)	-0.376 ± (0.04)	**0.177 ± (0.04)**	0.226 ± (0.1)	**0.060 ± (0.04)**
Avg. Odds. Diff. (Frac)	**-0.191 ± (0.02)**	0.197 ± (0.03)	**0.117 ± (0.1)**	0.201 ± (0.04)
Avg. Odds. Diff. (Rad)	-0.289 ± (0.07)	**0.138 ± (0.02)**	0.177 ± (0.1)	**0.044 ± (0.04)**
Dem. Parity Diff. (Frac)	**-0.109 ± (0.02)**	0.170 ± (0.04)	**0.083 ± (0.01)**	0.150 ± (0.05)
Dem. Parity Diff. (Rad)	-0.205 ± (0.10)	**0.118 ± (0.02)**	0.150 ± (0.04)	**0.045 ± (0.02)**
Bal. Acc. Drop (Frac)	**0.026**	**0.016**	0.012	**0.008**
Bal. Acc. Drop (Rad)	0.03	0.023	**0.009**	0.018

Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) when applying mitigation (through exponentiated gradient) for valvular heart disease diagnosis.

Prognosis

Arrhythmia

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	-0.526 ± (0.05)	**0.170 ± (0.05)**	**0.164 ± (0.02)**	**0.338 ± (0.08)**
Eq. Opp. Diff. (Rad)	**-0.380 ± (0.09)**	0.190 ± (0.03)	0.172 ± (0.03)	0.339 ± (0.09)
Avg. Odds. Diff. (Frac)	-0.510 ± (0.05)	**0.142 ± (0.04)**	**0.146 ± (0.02)**	0.293 ± (0.05)
Avg. Odds. Diff. (Rad)	**-0.350 ± (0.09)**	0.173 ± (0.03)	0.149 ± (0.02)	**0.281 ± (0.07)**
Dem. Parity Diff. (Frac)	-0.503 ± (0.03)	**0.151 ± (0.02)**	**0.143 ± (0.01)**	0.253 ± (0.03)
Dem. Parity Diff. (Rad)	**-0.345 ± (0.06)**	0.167 ± (0.03)	0.152 ± (0.02)	**0.231 ± (0.05)**

Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) for arrhythmia prognosis.

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	**-0.031 ± (0.01)**	**0.001 ± (0.00)**	**0.055 ± (0.02)**	**0.015 ± (0.01)**
Eq. Opp. Diff. (Rad)	-0.184 ± (0.09)	0.151 ± (0.04)	0.142 ± (0.03)	0.201 ± (0.08)
Avg. Odds. Diff. (Frac)	**0.002 ± (0.03)**	**0.001 ± (0.00)**	**0.032 ± (0.01)**	**0.001 ± (0.00)**
Avg. Odds. Diff. (Rad)	-0.167 ± (0.08)	0.128 ± (0.03)	0.112 ± (0.02)	0.137 ± (0.05)
Dem. Parity Diff. (Frac)	**-0.028 ± (0.01)**	**0.001 ± (0.00)**	**0.015 ± (0.01)**	**0.010 ± (0.01)**
Dem. Parity Diff. (Rad)	-0.161 ± (0.06)	0.130 ± (0.03)	0.108 ± (0.01)	0.081 ± (0.02)
Bal. Acc. Drop (Frac)	0.094	0.134	0.085	0.133
Bal. Acc. Drop (Rad)	**0.012**	**0.011**	**0.005**	**0.017**

Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) when applying mitigation (through exponentiated gradient) for arrhythmia prognosis.

Heart failure

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	-0.517 ± (0.18)	0.192 ± (0.07)	0.178 ± (0.07)	0.298 ± (0.11)
Eq. Opp. Diff. (Rad)	**-0.353 ± (0.05)**	**0.143 ± (0.02)**	**0.174 ± (0.04)**	**0.221 ± (0.07)**
Avg. Odds. Diff. (Frac)	-0.425 ± (0.14)	0.105 ± (0.09)	0.153 ± (0.04)	0.248 ± (0.09)
Avg. Odds. Diff. (Rad)	**-0.267 ± (0.07)**	**0.098 ± (0.05)**	**0.121 ± (0.05)**	**0.160 ± (0.07)**
Dem. Parity Diff. (Frac)	-0.401 ± (0.06)	0.161 ± (0.02)	**0.135 ± (0.02)**	0.219 ± (0.07)
Dem. Parity Diff. (Rad)	**-0.273 ± (0.02)**	**0.142 ± (0.03)**	0.150 ± (0.00)	**0.102 ± (0.06)**

Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) for heart failure prognosis.

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	**-0.116 ± (0.06)**	**0.110 ± (0.05)**	**0.023 ± (0.02)**	**0.088 ± (0.02)**
Eq. Opp. Diff. (Rad)	-0.350 ± (0.05)	0.140 ± (0.02)	0.171 ± (0.04)	0.212 ± (0.08)
Avg. Odds. Diff. (Frac)	**-0.016 ± (0.06)**	**0.001 ± (0.06)**	**0.011 ± (0.01)**	**0.020 ± (0.04)**
Avg. Odds. Diff. (Rad)	-0.266 ± (0.07)	0.090 ± (0.05)	0.116 ± (0.05)	0.155 ± (0.07)
Dem. Parity Diff. (Frac)	**-0.023 ± (0.02)**	**0.016 ± (0.02)**	**0.002 ± (0.00)**	**0.023 ± (0.01)**
Dem. Parity Diff. (Rad)	-0.271 ± (0.02)	0.135 ± (0.03)	0.148 ± (0.00)	0.100 ± (0.06)
Bal. Acc. Drop (Frac)	0.164	0.169	0.182	0.103
Bal. Acc. Drop (Rad)	**0.002**	**0.013**	**0.005**	**0.007**

Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) when applying mitigation (through exponentiated gradient) for heart failure prognosis.

Myocardial infarction

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	-0.949 ± (0.07)	**0.150 ± (0.03)**	**0.156 ± (0.04)**	0.258 ± (0.10)
Eq. Opp. Diff. (Rad)	**-0.648 ± (0.05)**	0.168 ± (0.03)	0.241 ± (0.04)	**0.205 ± (0.07)**
Avg. Odds. Diff. (Frac)	-0.901 ± (0.1)	**0.068 ± (0.03)**	**0.135 ± (0.04)**	0.222 ± (0.11)
Avg. Odds. Diff. (Rad)	**-0.614 ± (0.06)**	0.107 ± (0.02)	0.206 ± (0.04)	**0.161 ± (0.06)**
Dem. Parity Diff. (Frac)	-0.855 ± (0.13)	**0.155 ± (0.03)**	**0.143 ± (0.03)**	0.260 ± (0.10)
Dem. Parity Diff. (Rad)	**-0.585 ± (0.08)**	0.176 ± (0.03)	0.223 ± (0.01)	**0.169 ± (0.02)**

Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) for myocardial infarction prognosis.

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	**-0.164 ± (0.05)**	**0.107 ± (0.03)**	**0.102 ± (0.01)**	**0.068 ± (0.03)**
Eq. Opp. Diff. (Rad)	-0.366 ± (0.08)	0.154 ± (0.02)	0.202 ± (0.02)	0.148 ± (0.06)
Avg. Odds. Diff. (Frac)	**-0.111 ± (0.04)**	**0.040 ± (0.04)**	**0.083 ± (0.02)**	**0.053 ± (0.03)**
Avg. Odds. Diff. (Rad)	-0.308 ± (0.05)	0.098 ± (0.04)	0.179 ± (0.02)	0.093 ± (0.05)
Dem. Parity Diff. (Frac)	**-0.081 ± (0.04)**	**0.104 ± (0.02)**	**0.089 ± (0.03)**	**0.066 ± (0.02)**
Dem. Parity Diff. (Rad)	-0.256 ± (0.02)	0.163 ± (0.02)	0.188 ± (0.01)	0.091 ± (0.01)
Bal. Acc. Drop (Frac)	0.09	0.022	0.028	0.031
Bal. Acc. Drop (Rad)	**0.027**	**0.005**	**0.004**	**0.018**

Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) when applying mitigation (through exponentiated gradient) for myocardial infarction prognosis.

Stroke

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	**-0.177 ± (0.18)**	**0.123 ± (0.01)**	**0.135 ± (0.06)**	0.365 ± (0.1)
Eq. Opp. Diff. (Rad)	-0.408 ± (0.05)	0.210 ± (0.01)	0.337 ± (0.11)	**0.197 ± (0.04)**
Avg. Odds. Diff. (Frac)	**-0.134 ± (0.14)**	**0.072 ± (0.04)**	**0.022 ± (0.09)**	0.305 ± (0.07)
Avg. Odds. Diff. (Rad)	-0.360 ± (0.07)	0.179 ± (0.03)	0.252 ± (0.06)	**0.149 ± (0.03)**
Dem. Parity Diff. (Frac)	**-0.097 ± (0.1)**	**0.122 ± (0.01)**	**0.054 ± (0.04)**	0.281 ± (0.03)
Dem. Parity Diff. (Rad)	-0.408 ± (0.05)	0.189 ± (0.03)	0.226 ± (0.04)	**0.140 ± (0.03)**

Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) for stroke prognosis.

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	-0.171 ± (0.18)	0.100 ± (0.02)	0.100 ± (0.07)	0.225 ± (0.2)
Eq. Opp. Diff. (Rad)	**-0.064 ± (0.03)**	**0.052 ± (0.05)**	**0.096 ± (0.09)**	**0.115 ± (0.08)**
Avg. Odds. Diff. (Frac)	-0.108 ± (0.16)	0.059 ± (0.03)	**0.016 ± (0.07)**	0.127 ± (0.2)
Avg. Odds. Diff. (Rad)	**-0.024 ± (0.03)**	**0.010 ± (0.03)**	0.023 ± (0.06)	**0.059 ± (0.05)**
Dem. Parity Diff. (Frac)	-0.093 ± (0.10)	0.095 ± (0.02)	0.052 ± (0.04)	0.140 ± (0.12)
Dem. Parity Diff. (Rad)	**-0.015 ± (0.01)**	**0.006 ± (0.01)**	**0.002 ± (0.00)**	**0.018 ± (0.01)**
Bal. Acc. Drop (Frac)	**0.009**	**0.002**	**0.013**	**0.001**
Bal. Acc. Drop (Rad)	0.03	0.083	0.071	0.061

Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) when applying mitigation (through exponentiated gradient) for stroke prognosis.

Valvular heart disease

	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	**-0.361 ± (0.06)**	0.260 ± (0.03)	0.262 ± (0.10)	0.454 ± (0.10)
Eq. Opp. Diff. (Rad)	-0.452 ± (0.09)	**0.211 ± (0.05)**	**0.236 ± (0.09)**	**0.187 ± (0.03)**
Avg. Odds. Diff. (Frac)	**-0.326 ± (0.05)**	0.230 ± (0.01)	**0.158 ± (0.09)**	0.419 ± (0.09)
Avg. Odds. Diff. (Rad)	-0.371 ± (0.06)	**0.181 ± (0.03)**	0.206 ± (0.06)	**0.140 ± (0.02)**
Dem. Parity Diff. (Frac)	-0.300 ± (0.02)	0.205 ± (0.02)	**0.112 ± (0.01)**	0.399 ± (0.05)
Dem. Parity Diff. (Rad)	**-0.291 ± (0.04)**	**0.154 ± (0.02)**	0.176 ± (0.02)	**0.096 ± (0.03)**

Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) for valvular heart disease prognosis.

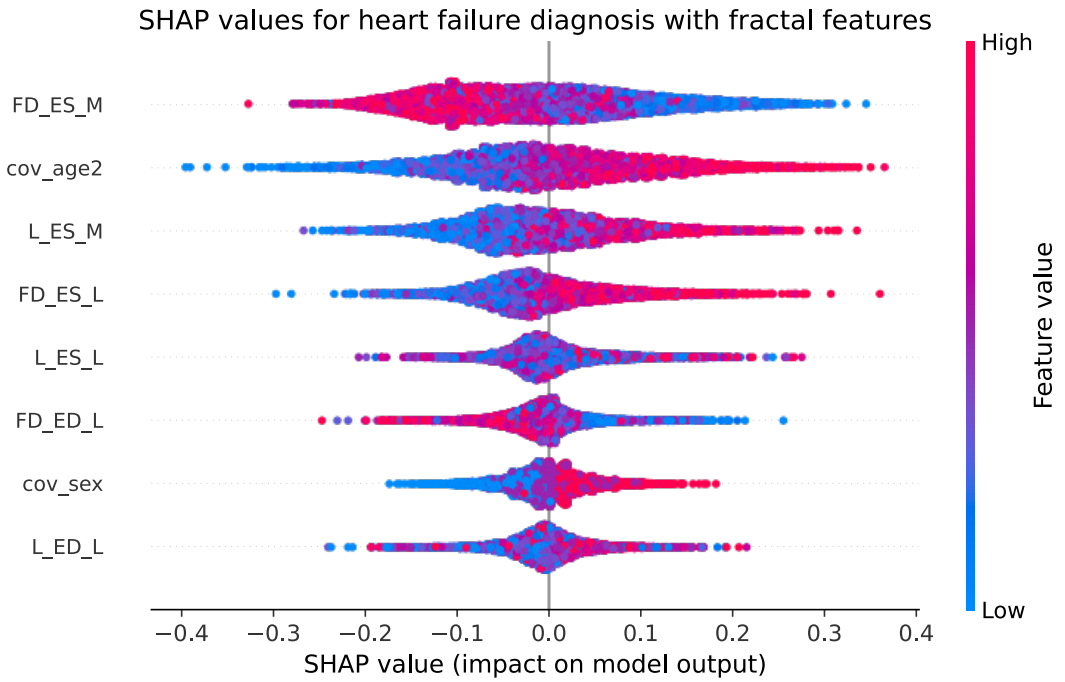
	Sex	Hypertension	Overweight or obese	Age
Eq. Opp. Diff. (Frac)	**-0.195 ± (0.05)**	0.200 ± (0.05)	0.253 ± (0.06)	0.194 ± (0.07)
Eq. Opp. Diff. (Rad)	-0.299 ± (0.15)	**0.172 ± (0.04)**	**0.170 ± (0.04)**	**0.132 ± (0.05)**
Avg. Odds. Diff. (Frac)	**-0.156 ± (0.03)**	0.180 ± (0.03)	**0.142 ± (0.03)**	0.154 ± (0.07)
Avg. Odds. Diff. (Rad)	-0.202 ± (0.1)	**0.154 ± (0.03)**	0.150 ± (0.03)	**0.099 ± (0.03)**
Dem. Parity Diff. (Frac)	-0.119 ± (0.02)	0.168 ± (0.02)	**0.093 ± (0.00)**	0.142 ± (0.03)
Dem. Parity Diff. (Rad)	**-0.107 ± (0.05)**	**0.141 ± (0.02)**	0.131 ± (0.01)	**0.068 ± (0.02)**
Bal. Acc. Drop (Frac)	0.031	0.019	0.013	**0.003**
Bal. Acc. Drop (Rad)	**0.004**	**0.009**	**0.003**	0.005

Mean and standard deviation of the fairness metrics obtained in the outer CV (testing) when applying mitigation (through exponentiated gradient) for valvular heart disease prognosis.

Annex E. Feature importance results for every component outcome.

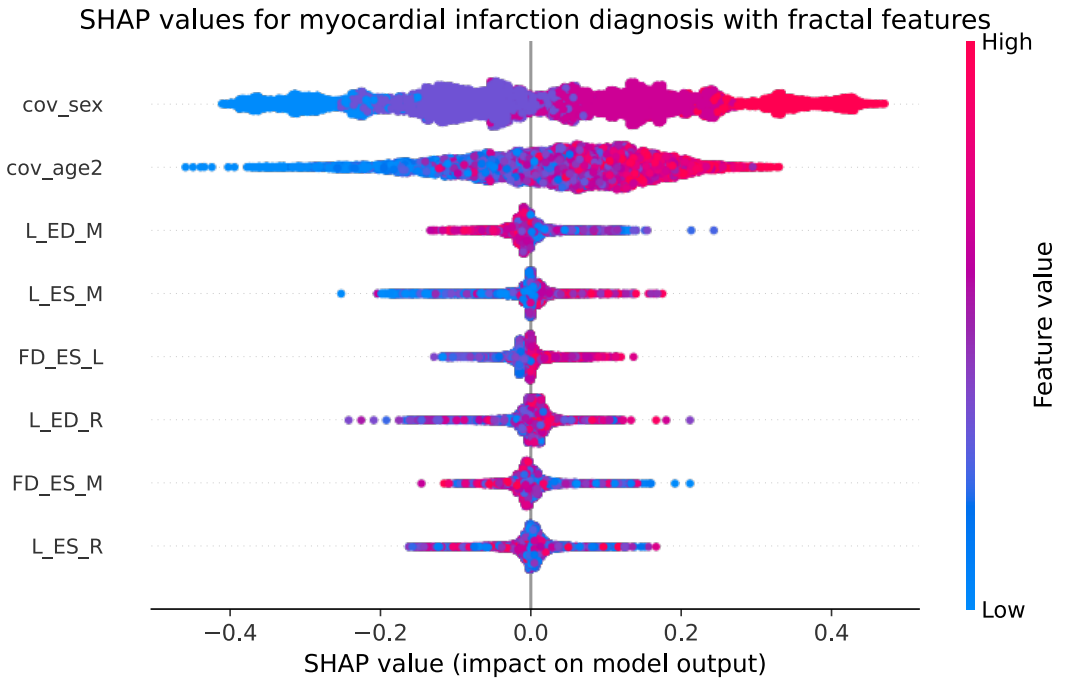
Diagnosis

Heart failure



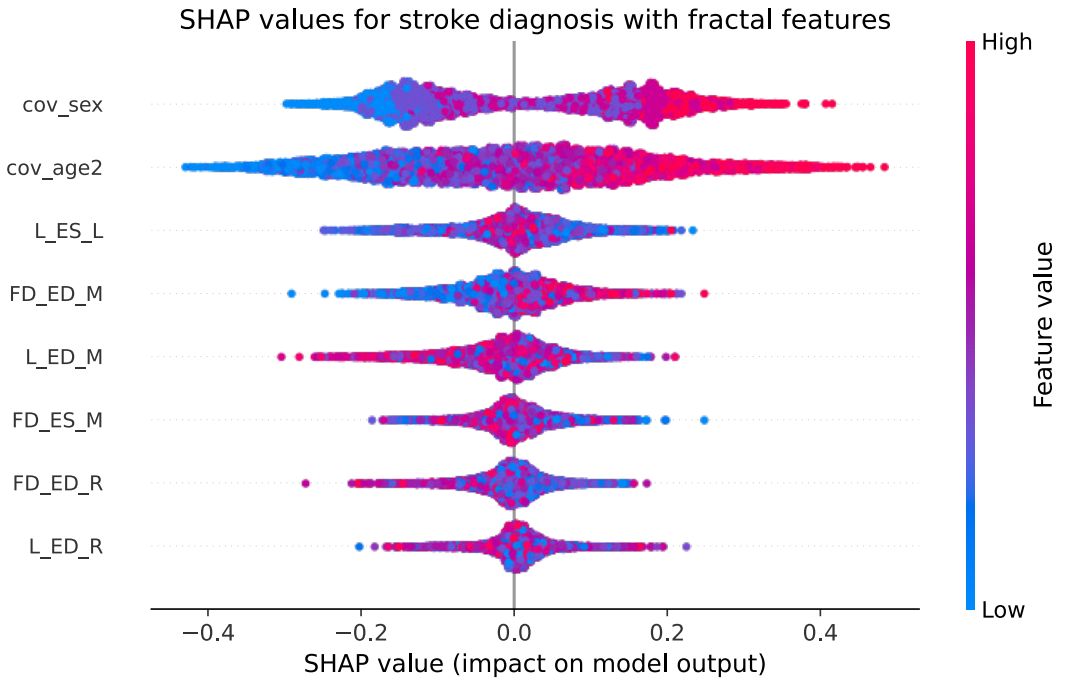
SHAP values of the top 8 feature of the best model for heart failure diagnosis using fractal-based features.

Myocardial infarction



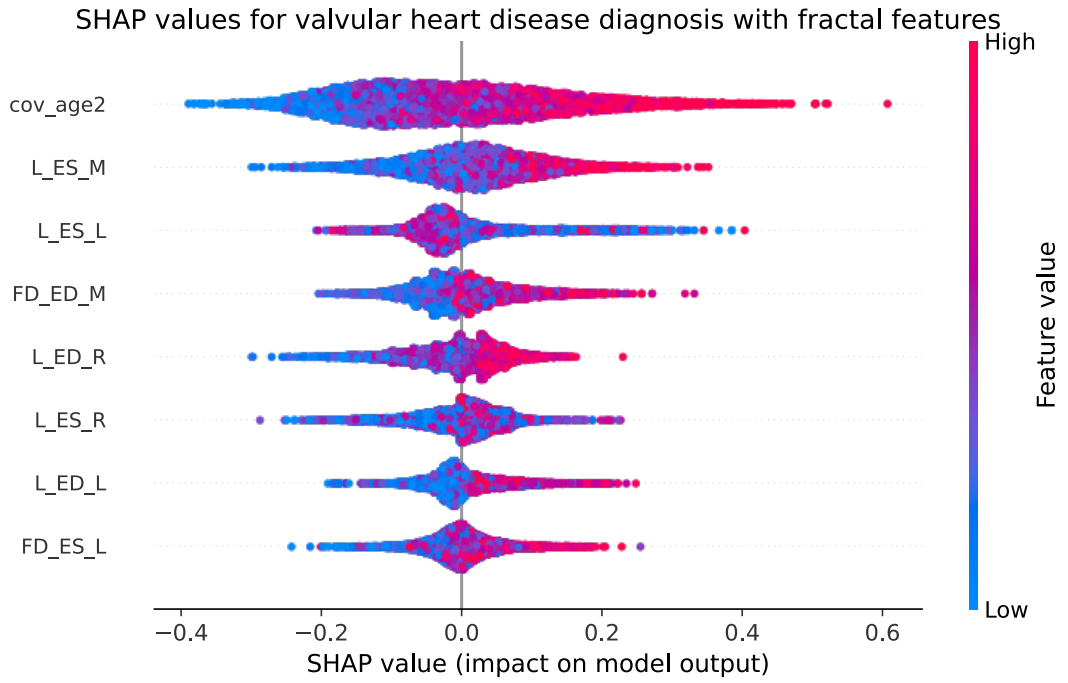
SHAP values of the top 8 feature of the best model for myocardial infarction diagnosis using fractal-based features.

Stroke



SHAP values of the top 8 feature of the best model for stroke diagnosis using fractal-based features.

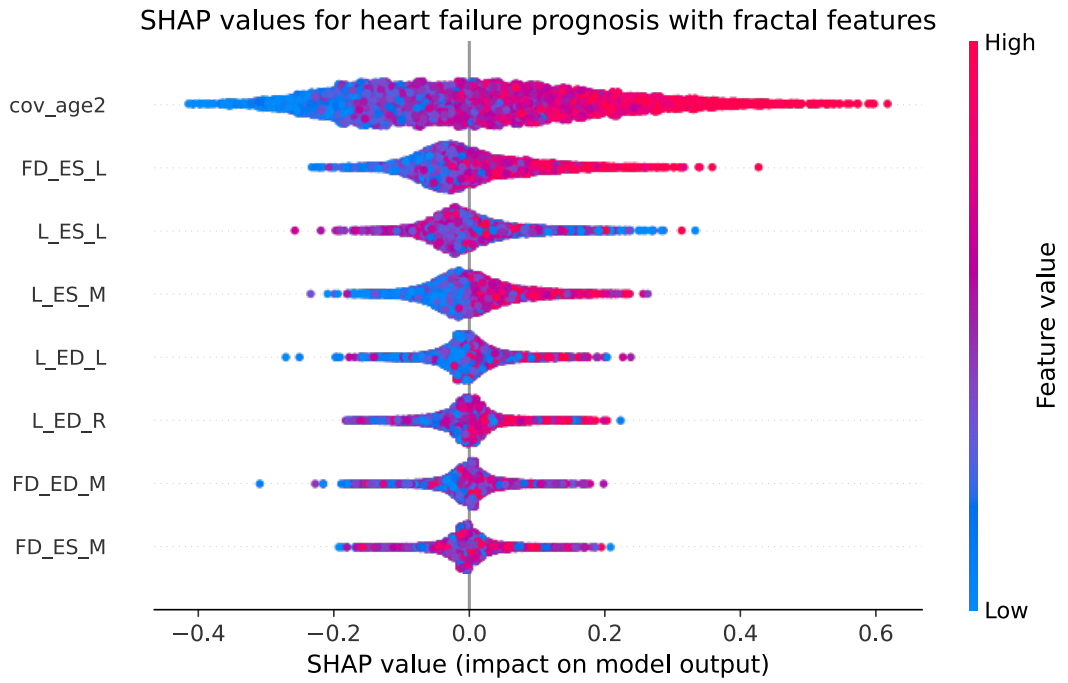
Valvular heart disease



SHAP values of the top 8 feature of the best model for valvular heart disease diagnosis using fractal-based features.

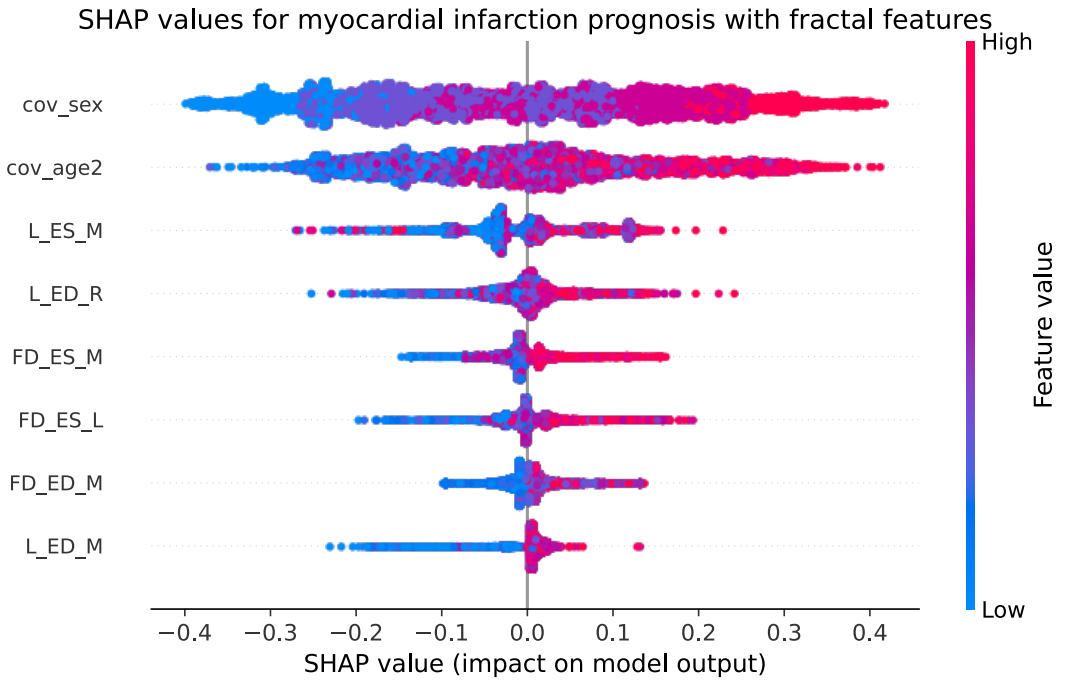
Prognosis

Heart failure



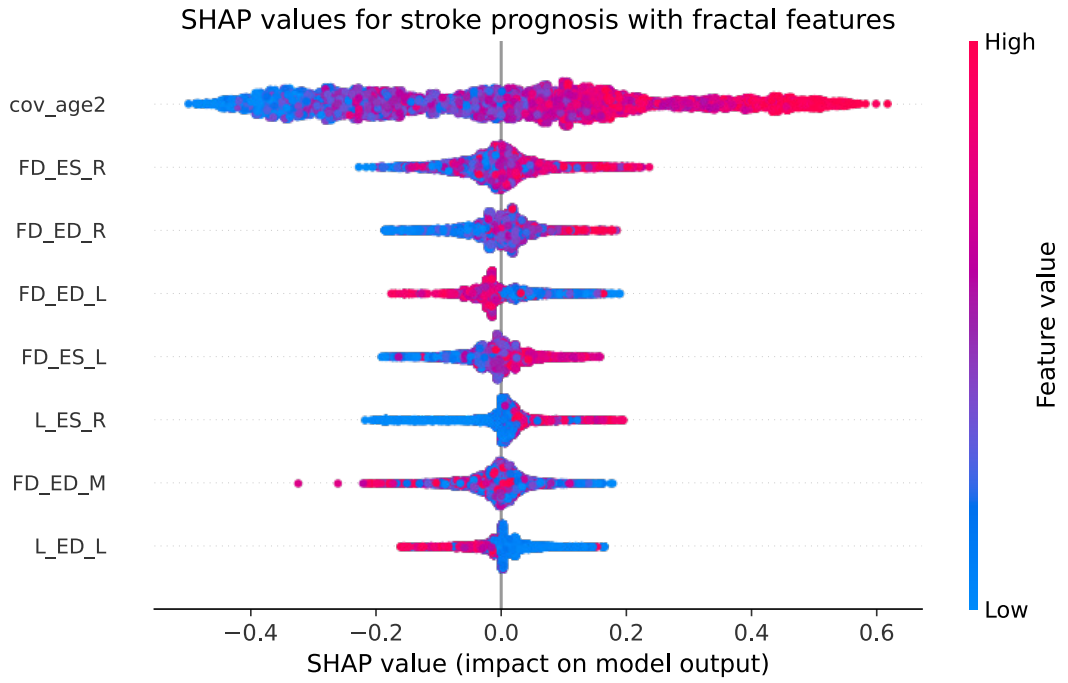
SHAP values of the top 8 feature of the best model for heart failure prognosis using fractal-based features.

Myocardial infarction



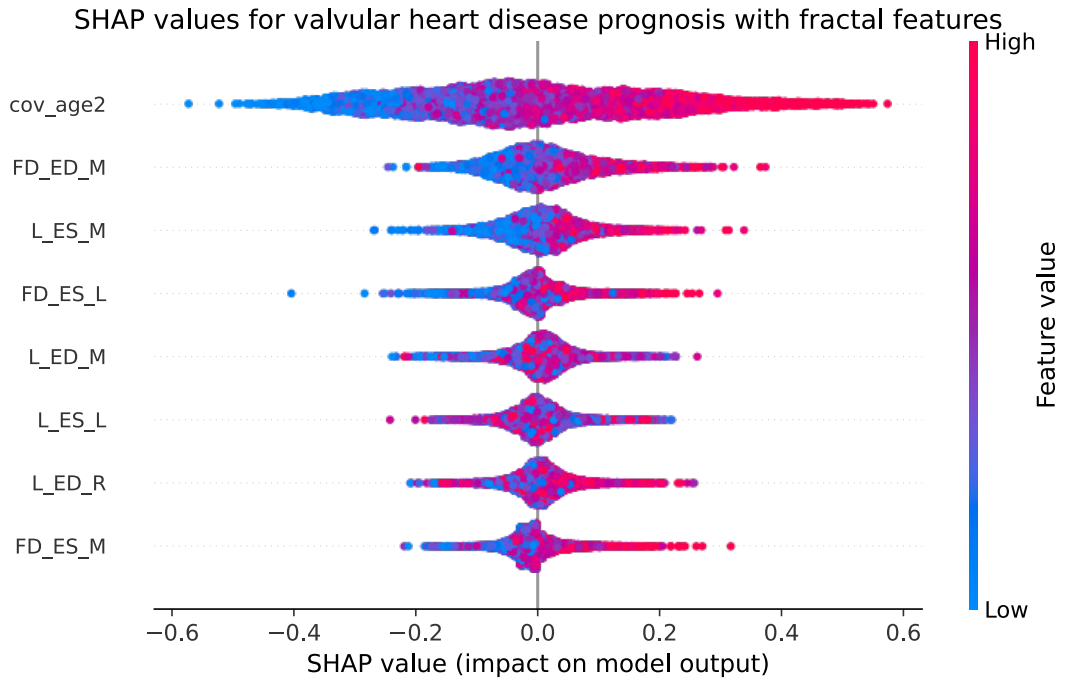
SHAP values of the top 8 feature of the best model for myocardial infarction prognosis using fractal-based features.

Stroke



SHAP values of the top 8 feature of the best model for stroke prognosis using fractal-based features.

Valvular heart disease



SHAP values of the top 8 feature of the best model for valvular heart disease prognosis using fractal-based features.

Annex F. Libraries

Library	Version
aif360	0.6.0
fairlearn	0.10.0
glob2	0.7
imbalanced-learn	0.10.1
ipykernel	6.29.3
ipython	8.22.2
ipywidgets	8.1.2
joblib	1.2.0
jupyter_client	8.6.1
jupyter_core	5.7.2
jupyterlab_widgets	3.0.10
lightgbm	4.3.0
matplotlib	3.7.1
matplotlib-inline	0.1.6
numpy	1.26.4
openpyxl	3.1.0
optuna	3.2.0
pandas	2.2.2
pip	24.0
scikit-learn	1.2.2
scipy	1.10.1
seaborn	0.13.2
shap	0.45.0
statsmodels	0.14.1
tensorboard	2.12.3
tensorboard-data-server	0.7.0
tornado	6.4
tqdm	4.66.2
xgboost	1.7.6

List of used libraries and their versions.