UNIVERSITY OF PALERMO

Department of Economics, Business and Statistics

Methodological Advances in Biological Complex Systems

Ph.D Candidate:

Vincenzo Martorana

Ph.D. Coordinator:

Prof. Vito M.R. Muggeo

DSEAS, UNIPA

Supervisor:

Prof. Michele Tumminello

DSEAS, UNIPA

Co-Supervisor:

Prof. Andrea Consiglio

DSEAS, UNIPA

Internship Supervisor:

Dr. Giorgia M. Bosi

MECENG Department, UCL

Company Supervisors:

Dr. Claudia Coronnello

Fondazione Ri.MED, Palermo

Prof. Gaetano Burriesci

Fondazione Ri.MED, Palermo

MECENG Department, UCL

Keywords: **Left Atrial Appendage; Unsupervised Detection System; Feature Extraction; Morphological Classification; Bioinformatics; Networks; Machine Learning**

XXXVI Ph.D. cycle - 2024







La borsa di dottorato è stata cofinanziata con risorse del
Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 (CCI 2014IT16M2OP005),
Fondo Sociale Europeo, Azione I.1 “Dottorati Innovativi con caratterizzazione Industriale”



UNIONE EUROPEA
Fondo Sociale Europeo





UNIVERSITY OF PALERMO
Department of Economics, Business and Statistics

Methodological Advances in Biological Complex Systems

Ph.D Candidate:
Vincenzo Martorana

Ph.D. Coordinator:
Prof. Vito M.R. Muggeo
DSEAS, UNIPA

Supervisor:
Prof. Michele Tumminello
DSEAS, UNIPA

Co-Supervisor:
Prof. Andrea Consiglio
DSEAS, UNIPA

Internship Supervisor:
Dr. Giorgia M. Bosi
MECENG Department, UCL

Company Supervisors:
Dr. Claudia Coronnello
Fondazione Ri.MED, Palermo

Prof. Gaetano Burriesci
Fondazione Ri.MED, Palermo
MECENG Department, UCL

Keywords: **Left Atrial Appendage; Unsupervised Detection System; Feature Extraction; Morphological Classification; Bioinformatics; Networks; Machine Learning**

XXXVI Ph.D. cycle - 2024

Contents

Introduction **9**

1 Left Atrial Appendage Detection and Feature Extraction **11**

1.1 Left Atrial Appendage: Role and Structure 12

1.1.1 Therapies for Clot Formation Risk Reduction 13

1.1.2 Left Atrial Appendage from Left Atrium separation . . . 15

1.1.3 Left Atrial Appendage Characteristics 17

1.2 Left Atrial Appendage Detection Method 20

1.2.1 Phase 1: Patients Data and Pre-Processing 21

1.2.2 Phase 2: LAA Detection 22

1.2.3 Phase 3: LAA Feature Extraction 29

1.3 Experimental Validation 32

1.4 Discussions 33

1.5 Conclusions 36

2 Advancement in Left Atrial Appendage Group Identity: A Feature-Based Clustering Approach **39**

2.1 Left Atrial Appendage Morphological Classification 40

2.1.1 State-of-the-art LAA Classification 41

2.1.2 Machine Learning Classification and Clustering Techniques 42

2.1.3 Network Analysis and Community Detection 52

2.2 LAA Data-Driven Group Identification 54

2.2.1 LAA Data-Driven Partitions 55

2.2.2 Clustering Validation 67

2.2.3 Partitions Comparison 70

2.3 Experimental Results 72

2.4 Discussions 81

2.5 Conclusions 84

3 MicroRNA and Messenger RNA Interaction Analysis 91
3.1 Biological Foundations and Experimental Techniques 93
3.1.1 Molecular Characterization of Key Biological Entities . . 93
3.1.2 Computational Approaches to miRNA Target Prediction 95
3.1.3 Gene Expression Analysis 96
3.1.4 Computational Methodologies in Gene Expression Regulation 97
3.1.5 Regulatory Networks of RNA 98
3.2 Methods 99
3.2.1 Data Preparation 100
3.2.2 Scenario Investigation 101
3.2.3 MicroRNA-mRNA bipartite network 102
3.2.4 Model evaluation 104
3.3 Experimental Validation 105
3.4 Discussion 108
3.5 Conclusion 109

Future Researches 111
3.6 Upgrade of LAA detection algorithm to extract more LAA features 111
3.7 Improvement of the Clustering Framework 112
3.8 Binding Probabilities Optimisation 113

Conclusions 115

List of Figures 119

List of Tables 121

Acronyms 123

Author contributions 127

A i

B xv

Acknowledgements xix

Bibliography xxi

Introduction

From intricate anatomical structures to molecular interactions, biological complex systems have witnessed remarkable attention in recent decades. Two areas of research within this domain are Left Atrial Appendage (LAA) morphological analysis and MicroRNAs (miRNAs) activity alterations. The methodological breakthroughs in the third millennium have not only deepened our understanding of these biological systems but have also paved the way for innovative approaches to address pressing medical challenges.

1. Left Atrial Appendage Morphological Analysis:
   Left Atrial Appendage (LAA) morphological analysis is a topic that has received particular attention in the context of atrial fibrillation (AF), the most common cardiac arrhythmia. LAA's morphology plays a pivotal role in thromboembolic risk assessment, as clots often form within its complex structure. Although there have been methodological advances in the field of lAA morphlogy analysis, the filed still suffers from clinical subjectivity.

2. MicroRNAs Activity Alterations:
   MicroRNAs (miRNAs) are regulators of gene expression. They extert their biological functions mainly by binding to target messenger RNA (mRNA) molecules and thus influencing various cellular processes. Understanding miRNA-mRNA interactions is essential in deciphering the molecular mechanisms underlying diseases like cancer, diabetes, and neurological disorders.

The thesis is structured into three chapters. The initial two chapters focus on the morphological analysis of the LAA, while the third chapter delves into alterations in miRNA activity.
We report a brief description of the thesis chapters:

1. **Left Atrial Appendage Detection and Feature Extraction**
   The first chapter regards the development of an automated system for detecting the LAA from triangulated surface mesh, overcoming the need for human intervention. As a biproduct of the detection process, the method characterises the left atrial appendage by computing specific morphological metrics (features).

2. **Left Atrial Appendage Morphological Classification**
   The second chapter aims to define a novel classification system of the LAA morphologies based on the quantitative features the method described in Chapter 1 computes. This approach significantly diminishes subjectivity in the analysis of the left atrial appendage, fostering greater consistency and standardization in the examination of its morphology.

3. **MicroRNA and Messenger RNA Interaction Analysis**
   The third chapter explores the intricate world of miRNA-mRNA interactions, emphasizing the importance of network analysis and competing models in deciphering these complex relationships.

Each of these chapters contributes to a comprehensive understanding of biological complex systems. They represent an evolution of research methodologies and their application in addressing complex medical challenges.

Chapter 1

Left Atrial Appendage Detection and Feature Extraction

The analysis presented in this chapter is part of the research submitted in the esteemed journal Computers in Biology and Medicine by the name *An Unsupervised Method to Detect the Left Atrial Appendage and Extract its Features.* The study of Left Atrial Appendage (LAA) morphology is an important branch of cardiovascular research. LAA is an outpouching of the left atrium (LA) where the left atrial thrombi responsible for thromboembolic events are most likely to be formed. LAA complex structure has a key role in the study of thromboembolic risk. It is fundamental for the identification of the best treatment in patients suffering from atrial fibrillation (AF).
This chapter introduces a comprehensive methodology for LAA detection, integrating multiple advanced algorithms. The outcome of this endeavour is a fully automated detection process that significantly reduces subjectivity in identifying the junction between the LAA and the LA. As a byproduct of the detection process, our method computes specific metrics designed to characterize the LAA. These advancements in results signify a substantial stride towards enhancing the morphological classification of the LAA, as elaborated in the subsequent chapter (Chapter 2).

Abstract

The morphological analysis of the left atrial appendage is crucial to selecting the most appropriate treatment for patients suffering from atrial fibrillation.

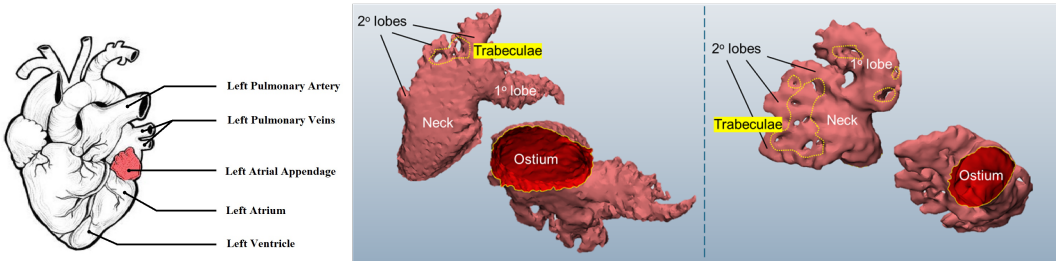
The analysis involves identifying the region that separates the left atrial appendage from the left atrium and calculating appendage features. Traditional methods encounter subjectivity issues in these two critical processes. We developed an automated system to tackle them. The system operates on a triangulated surface mesh from 3D reconstruction from clinical images. It exploits skeletonisation, distance analysis, affinity evaluation, and plane identification techniques for accurate appendage detection. Furthermore, it computes the left atrial appendage size and performs network analysis of the appendage skeleton to characterise the auricula using specific metrics. This approach offers a more objective, data-driven recognition and characterisation of the left atrial appendage.

1.1 Left Atrial Appendage: Role and Structure

Atrial Fibrillation (AF) stands as the most common type of cardiac arrhythmia, leading to chaotic and rapid heartbeats caused by faulty electrical impulses originating in the upper atria [89]. This disruption in heart rhythm affects blood flow, leading to fluid stagnation and clot formation. Such clot formation can give rise to thromboembolic events, including ischemic stroke [96] and vascular dementia [118]. Approximately 90% of left atrial thrombi form within the left atrial appendage (LAA) [111, 57].

The LAA begins developing during the embryonic stage of cardiac organogenesis, typically around the third to fifth week of gestation. Originating as a protrusion or outgrowth from the primitive left atrium, the LAA undergoes progressive growth and remodelling over subsequent weeks of gestation. Ultimately, the LAA assumes its definitive anatomical configuration within the fully formed heart [103, 124]. Upon maturation, LAA looks like an outpouching of the left atrium (LA) with a long, narrow, and sometimes wavy or hooked shape. It may feature a narrow junction and crenellated lumen [2, 144]. Its morphological complexity makes it prone to blood stasis, even in the absence of AF [2]. Figure 1.1 shows a graphical representation of the LAA position and some of its features.

Given the variability in LAA structure among individuals, clinicians and cardiologists must carefully evaluate its morphology to determine the most appropriate therapy for reducing the risk of clot formation.



(a) (b)

Figure 1.1: (a) Anatomical position of the Left Atrial Appendage. (b) Visual representation of two distinct LAA cases, highlighting the anatomical features such as the ostium, neck, primary and secondary lobes, and trabeculae.

1.1.1 Therapies for Clot Formation Risk Reduction

Reducing the risk of clot formation lies in choosing among three alternative therapies, each with its own set of adverse effects: oral anticoagulation, surgical exclusion, and percutaneous occlusion.

1. **Oral Anticoagulation Therapies**
   Anticoagulation drugs prevent the development of clots by targeting key components of the clotting cascade or directly inhibiting specific clotting factors in the blood [85]. In particular, by interfering with these processes, the therapy helps maintain blood flow and prevent the formation of blood clots within the circulatory system. International guidelines, such as the European Society of Cardiology (ESC) and the National Institute for Health and Care Excellence (NICE), recommend oral anticoagulation as a primary therapy for stroke prevention in AF patients at risk as it has been proven to be very effective [53, 135]. Anticoagulation therapy reduces the risk of clot formation, but it carries drawbacks, e.g. potential gastrointestinal bleeding and fatal intracranial haemorrhage. However, the risk of stroke in AF patients often outweighs the risk of bleeding associated with anticoagulants [63].

2. **Surgical Occlusion/Exclusion Techniques**
   Surgical techniques aim to reduce stroke risk by excising or excluding the LAA from the circulation. Principal techniques are excision [73], ligation [42], and obliteration [112]. Excision involves surgically removing the LAA, hence physically eliminating the potential source of blood

clots reducing the risk of stroke. The procedure involves the use of stapler, such as the Endo GIA II [55], which cut and seal tissues. Ligation involves closing off the opening of the appendage with ligatures, e.g. ENDOLOOP ligature [79], hence excluding it from circulation. Obliteration involves sealing the LAA through minimally invasive thoracoscopic procedures. This can be reached using epicardial clip like the AtriClip device [8]. It is a less invasive alternative to open-heart surgery and may be suitable for patients who are not candidates for traditional surgery. The abovementioned techniques can also be performed as part of a combined procedure during other cardiac surgeries, which may optimize patient outcomes and reduce the need for additional interventions.

3. **Percutaneous Occlusion Techniques**
   Percutaneous occlusion techniques represent a minimally invasive approach to reducing the risk of stroke in patients with AF. The Amplatzer AMULET Left Atrial Appendage Occluder (Abbott, Abbott Park, Illinois, USA) [115] and the WATCHMAN FLX LAA system (Boston Scientific, Marlborough, Massachusetts, USA) [102] devices are among the most commonly used. These devices are designed to be self-expanding and are inserted into the left atrial appendage (LAA) via a catheter inserted into a vein, typically in the leg.
   The procedure begins with accessing the femoral vein, through which the catheter is introduced. The catheter is guided through the venous system to the left atrium using advanced imaging techniques such as transesophageal echocardiography (TEE) or fluoroscopy. Once the catheter reaches the left atrium, the occlusion device is advanced through the catheter to the entrance of the LAA. When the device is in the correct position, it is deployed and automatically expands to fit the shape of the LAA. This process creates a physical barrier within the appendage, preventing the formation and release of thrombi. Over time, heart tissue grows around the device, integrating it into the heart and permanently occluding the appendage.
   This technique is particularly advantageous for patients who face a high risk of bleeding complications from traditional anticoagulant medications or who have contraindications to long-term anticoagulation therapy. It offers them a viable option for stroke prevention, significantly reducing the risk of thromboembolic events without the risks associated with an-

ticoagulants.

1.1.2 Left Atrial Appendage from Left Atrium separation

One of the primary challenges in the studies of LAA morphology is the lack of consistency in identifying the region that separates the LAA from the LA. In clinical practice, the appendage is often identified using the position and course of the left circumflex artery, but this practice is prone to subjectivity. Al-Saady et al. [2] pointed out the absence of a clearly marked junction that distinguishes the LAA from the LA and Ernst et al. [45] underlined there is no a clear anatomic boundary, while Tuccillo et al. [144] and Beigel et al. [9] highlighted a fairly well-defined narrow junction or orifice.

Manually identifying the region of separation between LAA and LA has been widely studied.
Ernst et al. [45] used LAA of cadavers in their study, where the appendage was separated from the LA identifying the orifice as the plane where there is a sudden change in width from LA to LAA.
Veinot et al. [148] highlighted the difference between the echocardiographic orifice and the anatomic orifice, manually identified from a point just inferior to the left superior pulmonary vein to the opposite side of the orifice.
Heist et al. [66] introduced the notation of neck to identify the site of reflection of the LAA with the surrounding LA. This site was then use to perform the analysis on the LAA.
Budge et al. [20] employed planar cardiac tomography (CTp), transesophageal echocardiography (TEE), and three-dimensional segmented CT reconstruction (CTsg) to address this issue. Specifically, they used multiple angles (sections) of TEE (CTp) to find the ones with the greatest orifice diameter, which aided in identifying the optimal line of demarcation between the LAA and LA. Additionally, they manually rotated the CTsg image to locate the LAA orifice based on its superior and inferior aspects relative to the pulmonary veins. This assessment was done by evaluating the curvature of the LA geometry and the continuation of the LAA wall.

Deep learning methodologies have recently been applied in numerous segmentation investigations of anatomical structures. Among these, certain studies have directed their focus towards detecting the LAA.
The approach of Jin et al. [72] first manually identified the region of interest (ROI) corresponding to the LAA in the computed tomographic angiography

(CTA) data . Then, a fully convolutional network (FCN), pre-trained on natural image segmentation, was fine-tuned for segmenting the 2D axial slices of the LAA images. Subsequently, the FCN is applied to each 2D slice individually to delineate the LAA boundaries accurately. Finally, a modified dense 3D conditional random field (CRF) refined the segmentations by incorporating spatial and contextual information.
Takayashiki et al. [140] used a U-Net architecture with five layers organised in an encoder-decoder structure. In the decoder section, the CT image is restored while retaining its positional information across layers to preserve spatial details. A 5x5 convolution filter in the encoder extracts features, with the parametric Rectified Linear Unit (ReLU) activation function applied afterwards. Down-sampling is performed using a 2x2 convolution filter with Stride 2 to capture a larger context and reduce computational complexity. Conversely, up-sampling is achieved with a 2x2 deconvolution filter with Stride 2, allowing the network to recover spatial details and produce segmentation results with the original image's resolution.
Chen et al. [29] employed a U-Net based architecture, applying Group Normalization (GN) at the end of each encoder and decoder block to enhance accuracy. Moreover, they implemented a negative example learning strategy to address the challenge of false positives by randomly sampling negative examples into the training batches.

Other authors developed automatic methodologies that do not require deep-learning procedures to identify and separate the LAA from the LA.
Masci et al. [97] developed an automatic LAA detection without human judgement. They used a 3D representation of the heart, resulting from in-house image segmentation algorithms for CT and magnetic resonance imaging (MRI) images, and utilised the shape diameter function (SDF) by Shapira et al. [131] to assign labels to each facet in the LA, pulmonary veins (PV) and LAA, based on iso-contours created by the SDF. The anatomical position of the PVs is used to recognise the appendage. The method is illustrated in Figure 1.2.
In a more recent approach, Lei et al. [86] employed an erosion-dilation image processing technique to identify LAAs with intricate and trabeculated structures. Their method utilises reconstructed 3D mesh from CT and TEE scans of the heart, where they apply the Hough transform to delineate the position of the aorta and subsequently remove it along with both left ventricle (LV) tissues from the heart. A corrosion operation is applied to the remaining heart

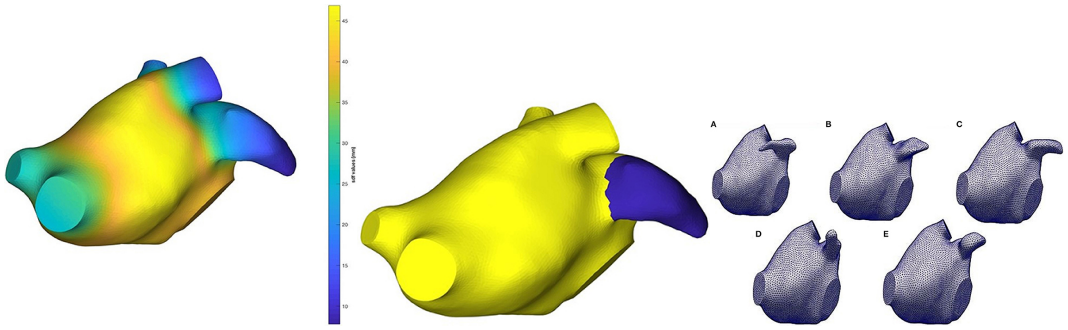
part, isolating only the left atrium. Finally, the isolated left atrium undergoes dilation using a dilate operation, and this dilated structure is then subtracted from the original mesh. The method is illustrated in Figure 1.2.

1.1.3 Left Atrial Appendage Characteristics

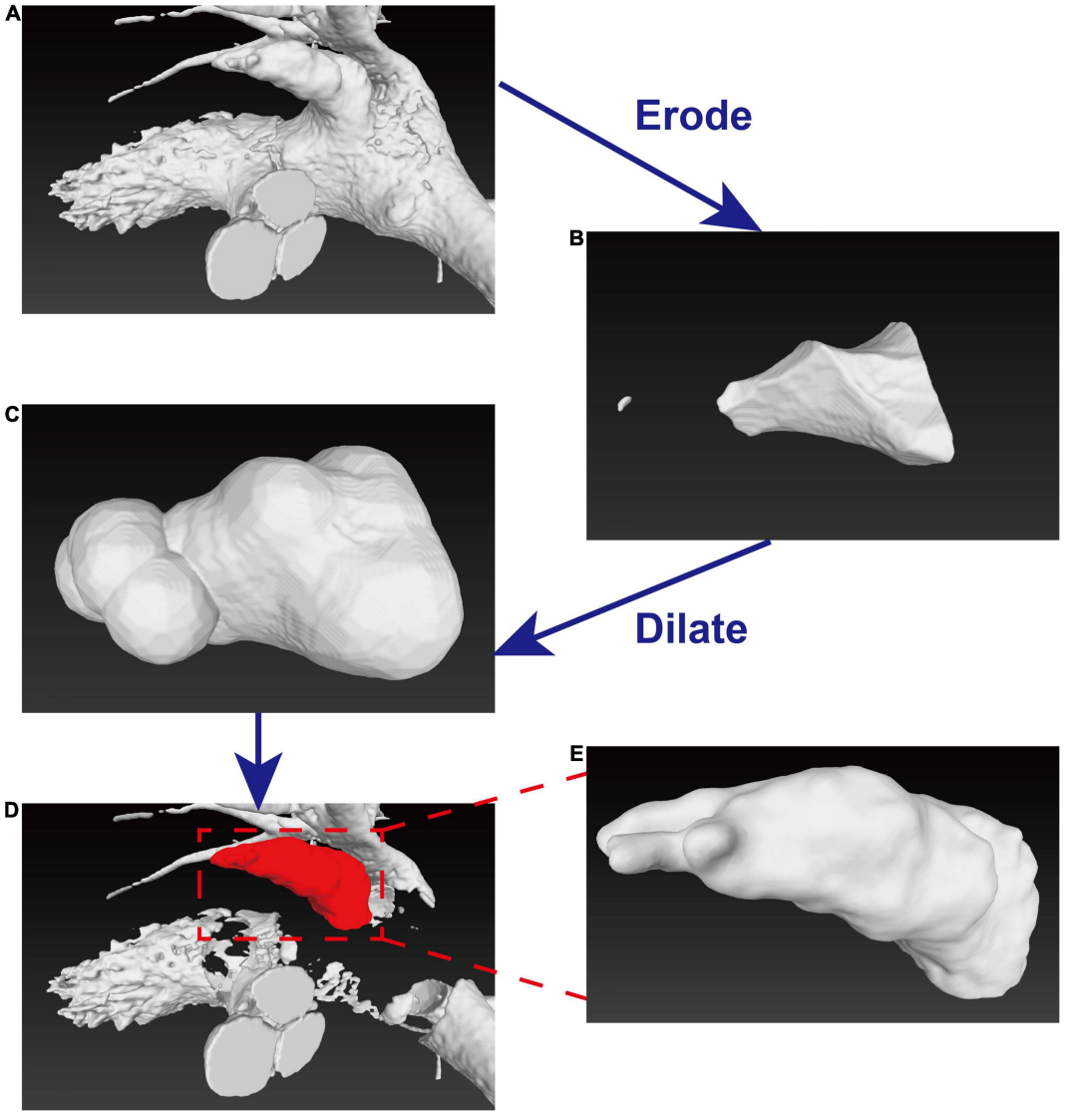
The structural complexity of the LAA plays a pivotal role in the pathogenesis of thrombus formation, with its anatomical characteristics contributing to the risk. Many studies have investigated this relationship. For instance, Ernst et al. [45] found that patients with atrial fibrillation (AF), left ventricular hypertrophy, and myocardial scars tend to have larger LAA volumes and ostium sizes, thus aiding in the formation of clots. Shirani and Alaeddini [132] observed that in patients with dilated appendages and chronic AF, the volume fraction of pectinate muscles is lower (i.e. the proportion of the total LAA volume that is occupied by the pectinate muscles), and the LAA surface appears smoother, consequently promoting the transition of clots. Additionally, Yamamoto et al. [161] noted that numerous lobes in the LAA may lead to blood stasis, potentially contributing to clot formation.

The LAA ostium, also LAA orifice, stands for the opening through which the appendage communicates with the LA and it is often a measured feature. Ernst et al. [45] measured the minimum and maximum diameters of the flat surface of the casts. Veinot et al. [148] computed the width of the orifice as the maximum diameter. Heist et al. [66] used the neck as a proxy of the orifice and then measured it with minimum and maximum diameter. Budge et al. [20] measured the minimum and maximum diameters and Yamamoto et al. [161] included the computation of orifice area and Li et al. [88] the computation of orifice perimeter.

Tuccillo et al. [144] highlighted the tubular structure of the LAA. Ernst et al. [45] measured its shape through the perpendicular distance between the orifice plane and the end of the principal course and the maximum distance in the plane perpendicular to the orifice plane. Moreover, they measured the LAA volume, computing the volume of the cast. Like them, Veinot et al. [148] measured the perpendicular distance and the appendage length as the curvilinear distance from the anatomical orifice to the tip. They also measured the width as the widest external distance. Shirani and Alaeddini [132] computed the volume of transplanted LAAs by filling them with normal saline. Finally, they measured the luminal surface and pectinate muscle areas after bisecting



(a) (b) (c)



(d)

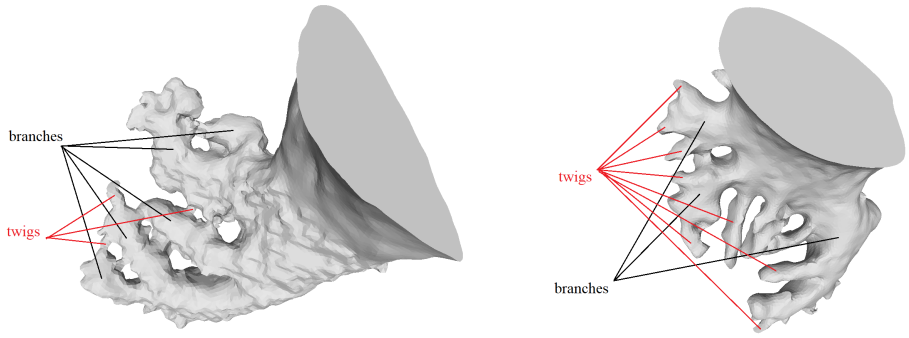
Figure 1.2: The first row of the figure shows the automatic detection developed by Masci et al. [97], where the shape is coloured using the SDF (a), and the appendage detected (b). (c) shows all the data utilised in their work. (d) shows the erosion-dilation detection algorithm of Lei et al. [86].

LAAs. Heist et al. [66] used 3D MRI models to compute the LAA and the depth, computed as the distance from the neck centre to the LAA greatest linear extent. Budge et al. [20], Yamamoto et al. [161], Li et al. [88] measured LAA volume, the length and width of the main lobe, where the LAA depth is defined as the longest straight line drawn within the main lobe from the midpoint of the orifice.

The narrow shape is another feature that characterises the LAA [144]. The principal course was categorised into five groups by Ernst et al. [45] according to the angle and the number of direction changes. In particular, they categorised the main course in three angle possibilities: straight if LAA present no angle smaller than 160°; slightly bent if it presents at least one change within the range 100°-160° but none smaller than 100°; and extremely bent if LAA has at least one change below 100°. Two direction change categories are allowed: slightly spiral if a virtual plane could divide the LAA into two almost symmetrical halves, and extremely spiral otherwise. The plausible scenarios are straight, slightly bent and slightly spiral, slightly bent and extremely spiral, extremely bent and slightly spiral, and extremely bent and extremely spiral. The first bend angle was also computed by Li et al. [88].

Ramifications of the principal course are complex structures, too. Ernst et al. [45] followed the botanical system defining branches, twigs, and fine structures not based on their length but on their orifice area and starting position. Branches were defined as ramifications with an orifice area larger than 10 mm^2, twigs as those with an orifice area ranging from 1 mm^2 to 10 mm^2, and fine structures as those with an orifice area smaller than 1 mm^2. While the branches are ramified of the principal course only, the twigs can ramify from both the principal axis and branches. Fine structures can ramify from the principal axis, branches and twigs (Figure 1.3 [1]). The number of branches was categorised into three groups (2 to 3, 4 to 5, more than 5), number of twigs was categorised into three groups too (less than 20, 20 to 40, more than 40), and number of fine structures into four classes (none or hardly any, occasional, frequent, dense). Veinot et al. [148] introduced the lobe notation by defining it as a visible protrusion from the main body but lying on a different plane and capable of admitting a 2 mm probe. According to this definition, the number of lobes was also measured by Heist et al. [66], Budge et al. [20], Yamamoto

[1] Recognising twigs using a triangulated surface mesh is not a simple task. The challenge increases and becomes almost impossible when identifying fine structures.



(a) (b)

Figure 1.3: Example of branches and twigs recognition in the analised data.

et al. [161].

Finally, the crenellation of LAA is another factor highlighted as correlated to clot formation risk. Tuccillo et al. [144] described the LAA as having a crenellated lumen.

1.2 Left Atrial Appendage Detection Method

The LAA detection method presented in this section is schematised in Figure 1.4 and structured into four distinct phases. In particular, Phases 1 to 3 are described in this chapter, while Phase 4 is extensively details in the following chapter (Chapter 2). The input data consists of 3D models derived from medical images containing the left atrium, the appendage, and the pulmonary veins (LA-LAA-PV), with the PVs truncated close to the LA. During Phase 1, the data are collected and pre-processed. Phase 2 focuses on detecting the left atrial appendage in an unsupervised manner. It employs several algorithms, including skeletonisation, distance computation, threshold detection, affinity evaluation, complexity analysis, and plane identification. These techniques collectively allow to identify the appendage accurately. In Phase 3, various features related to the shape and skeleton of the LAA are computed. This involves evaluating the geometric characteristics of the LAA shape, such as volume, ostium perimeter, and ostium area.

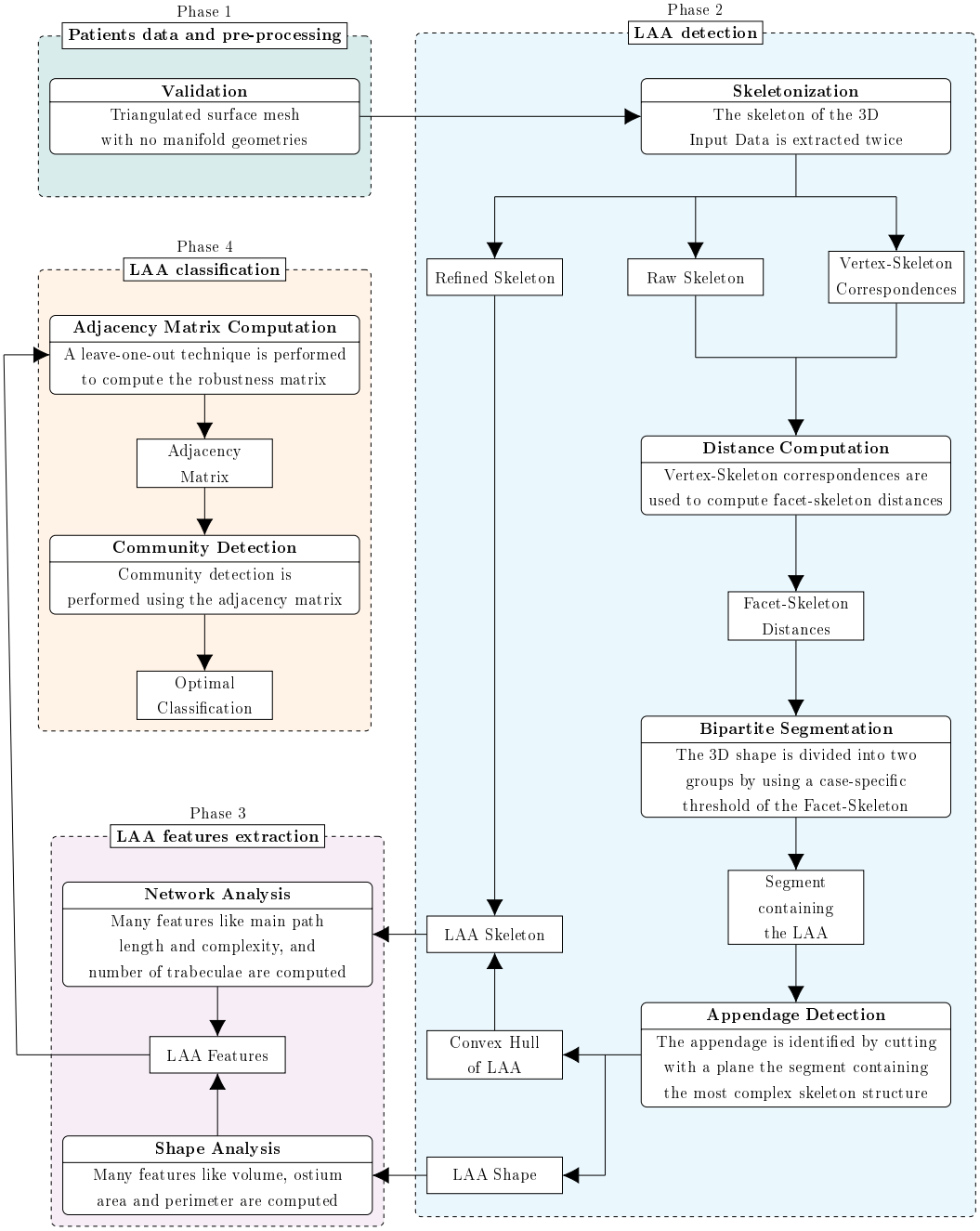


Figure 1.4: Flow chart of the left atrial appendage detection method.

1.2.1 Phase 1: Patients Data and Pre-Processing

The method presented in this work operates on 3D shapes of the LA and LAA, which are derived from routinely acquired clinical computerised tomography (CT) images using the commercial software Mimics (Materialise, Leuven, Belgium). Although the clinical CT image segmentation is not within the scope of this study, it constitutes a pre-processing phase.

The clinical CT images present row and column spacing of 0.488 mm and a slice thickness of 0.625 mm. In other words, the distance between the centres of two adjacent pixels in a CT image equals 0.488 mm, and the distance between two subsequent CT images equals 0.625 mm. Consequently, this methodology can capture the trabeculae, as their dimension ranges in $[0.1, 12]$ mm^2. Our previous experience has shown that details smaller than 0.5 mm have a negligible impact on the structural behaviour of models [105, 106]. Therefore, the CT images provide sufficient resolution for adequately reconstructing the LAA inner structure.
The clinical CT images were segmented in the post-processing program Mimics to derive the 3D anatomical shape of LAA and LA, following a common segmentation method [18]. The first step involves applying a manual threshold to create 2D masks of the region of interest based on grey values (in these cases, those highlighted by contrast medium). This is followed by a cropping process and a region-growing algorithm to eliminate disconnected parts. Manual editing ensures the inclusion of all relevant anatomical structures in the segmentation. Finally, a 3D virtual object is created, followed by a smoothing and remeshing step to avoid faulty mesh triangles.

The manual process for such segmentation could impact the result of the 3D model [27]; however, it currently outdoes the available automatic segmentation algorithms, which work for relatively simpler structures, such as the LV and the LA, but are not applicable to the complex and highly variable shape of the LAA, which is rich in small details such as trabeculations.

As the pre-processing phase is not within the scope of this work, we used the procedure mentioned above to create the 3D model because it provides a detailed triangulated surface mesh. Manual or not, any other technique can also be used in this scope if it provides a triangulated surface mesh, the unique input of the automated detection system.

The obtained triangulated surface mesh can matemathically be defined as $\mathcal{M} = (V, E)$, where $V = \{v_1, v_2, \ldots, v_n\}$ represents the set of vertices, and $E = \{e_{ij}\}$ represents the set of edges with $i \neq j$ and $i, j \in V$.

1.2.2 Phase 2: LAA Detection

The second phase of the proposed method is illustrated in Phase 2 of Figure 1.4, with a graphical representation provided in Figure 1.5. The input to this phase is a triangulated surface mesh with no-manifold geometries, as shown in

Figure 1.5a. The phase consists of four distinct steps: Skeletonisation, Distance Computation, Bipartite Segmentation, and Appendage Detection. Skeletonisation aims to capture the intricate structure of the model and describe it with a spatial graph composed of vertices and edges. Distance Computation characterises the mesh faces with a higher measure as the surface is closest to the skeleton. Bipartite Segmentation divides the surface into two groups according to a patient-specific threshold. Finally, Appendage Detection allows the system to identify and cut the appendage from the atrium correctly.

Skeletonisation

The step employs the well-established algorithm developed by Tagliasacchi et al. [139] to extract the skeleton of the mesh, denoted as $\mathcal{S}_{\mathcal{M}}$. The process involves iteratively contracting the input mesh using the mean curvature flow (MCF) until the shape volume vanishes. Each iteration follows the discretised MCF, which utilises the cotangent Laplacian operator L as an implicit scheme. To solve the problem effectively, the algorithm minimise the energy function given by:

$$E = E_{smooth} + E_{velocity} + E_{medial}, \tag{1.1}$$

where

- $E_{smooth} = \|W_L L V^{t+1}\|^2$ is the term used to smooth the skeleton;
- $E_{velocity} = \sum_i w_H \left\|v_i^{t+1} - v_i^t\right\|^2$ is the term used as a quality-time trade-off;
- $E_{medial} = \sum_i w_M \left\|v_i^{t+1} - \mu(v_i^t)\right\|^2$ is the term to lie the skeleton on the medial axis.

The parameters w_L, w_H, and w_M are chosen such that their product is equal to 1, ensuring a balanced effect. Notably, employing MCF promotes surface mesh contraction in the direction of the largest curvature, leading to increased anisotropy and the formation of badly shaped triangles. To mitigate these issues, edges with lengths below a specified distance ($\epsilon = 0.002$) are collapsed, and faces with one angle greater than a specified value ($\alpha = 110°$) are split. Additional steps involve iteratively collapsing the shortest edge on a vertex with two neighbouring faces. The result of the algorithm is a medially centred curve skeleton that preserves the model topology and can be represented as a spatial graph, i.e. a set of vertices linked through edges. The nodes with

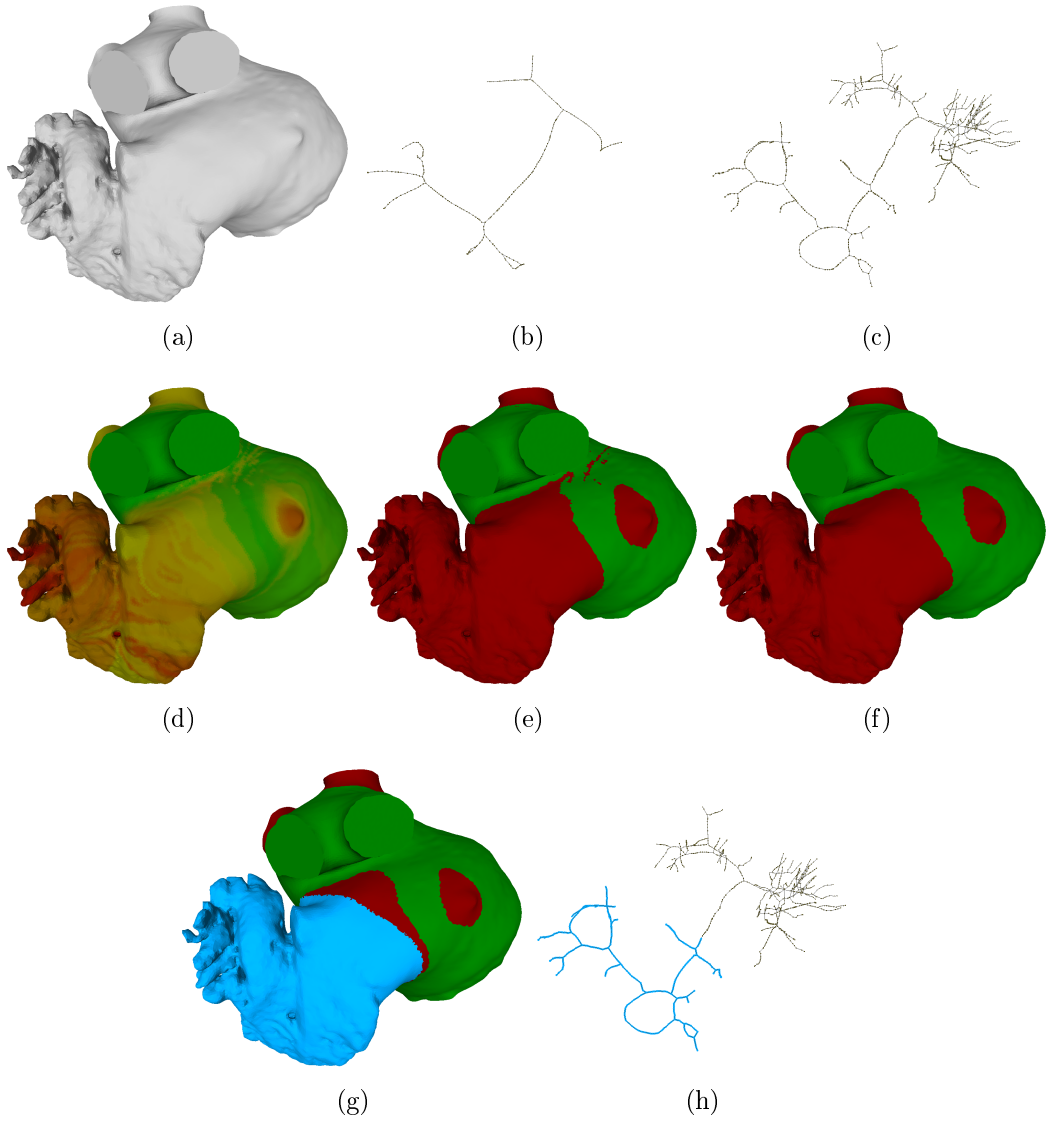


Figure 1.5: Given a triangulated surface mesh (a), two skeletons are extracted: a simple one (b) and a complex one (c). The facet-skeleton distances are computed using the simple skeleton, and the distances are used to map the mesh surface (d). The shape is divided into two groups (red and green) using a threshold measure (e). The borders of the groups are smoothed by evaluating the membership affinity of the faces (f). The LAA (cyan) is detected by identifying the connected component of the red group with the most complex inward skeleton, and a plane is used to cut it with a net cut (g). The LAA's skeleton (cyan) is obtained by cutting the complex skeleton with the plane (h).

degree 1 are terminal nodes, the ones with degree 2 serve as descriptors of the torturous path of the skeleton, and nodes with degrees higher than 2 represent

branching nodes. During the process, the algorithm also keeps track of the correspondences between the input mesh vertices v_i and the skeleton vertices $s_i \in \mathcal{S}_\mathcal{M}$. The skeletonisation process is performed twice with different parameter values. The first skeleton (Figure 1.5b) is generated using the default value of parameters proposed by Tagliasacchi et al. [139]: $w_H = 0.1$ and $w_M = 0.2$. It is a raw approximation of the skeleton, and as it contains only a few points, it is a suitable candidate for computing the facet-skeleton distances. The correspondences between the skeleton points and the shape vertices are stored in a set of pairs $P = \{(s_j, v_i)\}$. The second skeleton (Figure 1.5c), more detailed than the former, is used after the LAA detection to identify the LAA features. This is computed using different parameter values: $w_H = 0.5$ and $w_M = 2.0$. Although these parameter values penalised time, they allowed for a more detailed skeleton. In particular, reducing the smoother parameter w_L from 50 to 1 allows us to obtain a curve skeleton that is less smooth than the previous. Moreover, the rise of the medial parameter w_M from 0.2 to 2 allows the skeleton to follow the fine structures of the appendage very well and capture the trabeculations. It will be used after the LAA detection to compute some features of its structure.

Distance Computation

The method uses the correspondences set among shape vertices $v_i \in V$ and skeleton vertices $s_j \in \mathcal{S}_\mathcal{M}$ found in the skeletonisation step to compute the distance between the faces and the skeleton. For each pair $(s_j, v_i) \in P$, the Euclidean distance between v_i and the corresponding skeleton point s_j is computed and denoted as $d(v_i, s_j)$. Any shape vertices linked to a skeleton vertex are grouped together and characterised by a measure equal to the average distance. For a skeleton vertex s_j, with m linked vertices v_i, the averaged distance is obtained as follows

$$d'(v_i, \mathcal{S}_\mathcal{M}) = \frac{1}{m} \sum_{i=1}^{m} d(v_i, s_j). \tag{1.2}$$

The new measure $d'(v_i, \mathcal{S}_\mathcal{M})$ is then assigned as the sphere radius for the m corresponding vertices v_i. Let f_{ijk} represent the facet composed of three vertices v_i, v_j, and v_k, such that the three edges of the facet are e_{ij}, e_{jk}, and e_{ki}. The distance between facet f_{ijk} and the skeleton is the average of the three

vertex-skeleton distances found in (1.2) [2]:

$$d(f_{ijk}, \mathcal{S}_{\mathcal{M}}) = \frac{d'(v_i, \mathcal{S}_{\mathcal{M}}) + d'(v_j, \mathcal{S}_{\mathcal{M}}) + d'(v_k, \mathcal{S}_{\mathcal{M}})}{3}. \tag{1.3}$$

The resulting distances are visualised as a color map function of the surface in Figure 1.5d.

Bipartite Segmentation and Affinity Evaluation

While tracing the contour of the LAA, starting from the appendage apex, where the shortest distances to the skeleton are evident, these distances progressively increase as one approaches the atrium. To accurately locate the LAA ostium, our method incorporates a patient-specific distance threshold. This threshold categorizes the faces into two groups, a concept inspired by the segmentation method proposed by Otsu [113] for grey-level picture segmentation. The determination of the threshold is achieved through an unsupervised and non-parametric optimization process. The faces are divided into two groups, C_0 and C_1, by a break-point at a distance k, such that C_0 comprises faces with $d(f_{ijk}, \mathcal{S}_{\mathcal{M}}) \leq k$, and C_1 contains faces with $d(f_{ijk}, \mathcal{S}_{\mathcal{M}}) > k$. The optimal threshold k^* is determined by solving the following optimisation problem

$$\max_k \sigma_B^2(k), \tag{1.4}$$

where $\sigma_B^2(k)$ is the between-class variance for the groups identified by k and computed as $\sigma_B^2(k) = n_0(\mu_0 - \mu)^2 + n_1(\mu_1 - \mu)^2$.

After identifying the threshold, the shape is divided into two groups C_0^* and C_1^*, where C_0^* contains the component that corresponds to the LAA (Figure 1.5e). However, due to border irregularities, a refinement procedure is required to smoothen the borders and remove small marks. The refinement involves iteratively evaluating the affinity of each facet and its neighbours. For each facet $f \in C_0^*$, we change its membership if more than 50% of the first neighbours belong to C_1^*. This evaluation is repeated until no face changes its membership. The same procedure is applied to the faces $f \in C_1^*$. The 1st-order evaluation is needed to address little irregularities like holes and sharp borders (Figure 1.6). To achieve smooth borders while maintaining the separation of different parts within the same group, the affinity evaluation for high-order neighbourhoods

[2] Possible alternatives might have involved edge-edge distances or facet-edge distances between the shape and the skeleton. However, we have found that vertex-vertex distances yield better results than other distances.

is performed only for faces belonging to C_0^*. In this work, the affinity for faces $f \in C_0^*$ is evaluated up to the 9th-order neighbours, striking a quality-time trade-off. If at least one facet changes its group during the evaluation, the procedure is started from the first order. To prevent infinite loops caused by continuous changes in membership, the method keeps track of the number of times a facet changes its membership and halts the evaluation if it reaches 100. The result of the refinement procedure is shown in Figure 1.5f, where the borders are smoothened, and there are no more irregularities.

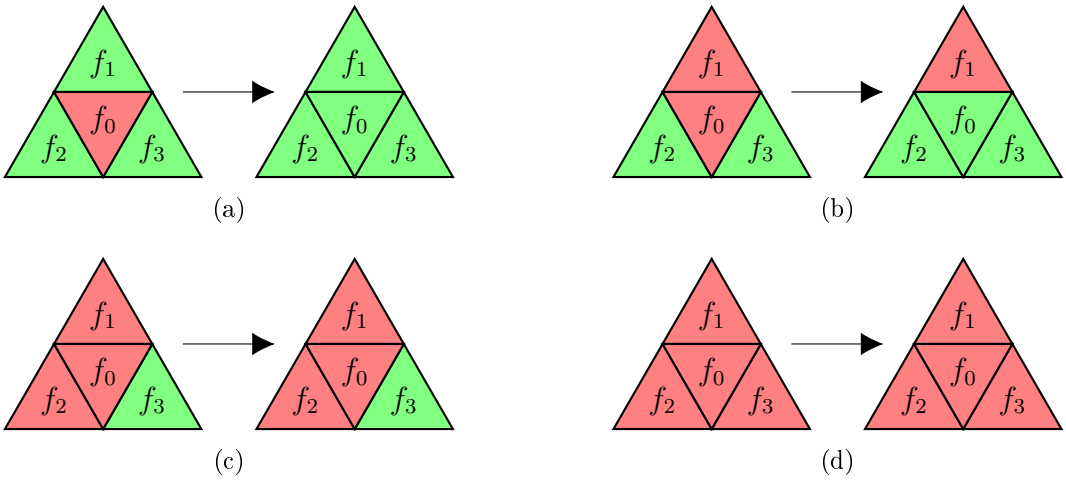


Figure 1.6: Affinity refinement used to fill little irregularities. The candidate facet f_0 belonging to a given group changes its membership (cases a and b) if it has more than 50% of its neighbours belonging to the other group.

Plane Identification and LAA from LA Separation

The final step involves detecting the LAA by analysing the density of the refined skeleton parts in each group of C_0^*. The appendage is identified as the group in C_0^* that exhibits the higher skeleton density. To measure the skeleton density of each group, the method counts the number of skeleton points encompassed within the convex hull of the group. The use of the convex hull mesh is preferred over the group mesh itself since the skeleton characterising the trabeculae may fall outside the mesh but remains within the convex envelope.

After detecting the LAA, a plane cut is performed along the border of the mesh to better emulate results typically obtained from clinical imaging techniques. The plane is identified using the first and second components of the Principal

Component Analysis (PCA) applied to the vertices belonging to the mesh border. The procedure removes the border vertices and correspondent faces that fall on the opposite side of the appendage, with respect to the identified plane. The plane computation is iteratively performed until two successive iterations return the same plane. The result of the detected appendage is displayed in Figure 1.5g.

The adequacy of the mesh cut is determined based on whether it creates only one hole, representing the LAA ostium. If multiple holes are generated due to a cut along the trabeculae (Figure 1.7a and 1.7b), it indicates that the appendage is cut too short and lacks a neck. In such cases, the procedure returns to the skeletonisation step and repeats the entire process.

Following the mesh cut, an ellipse is fitted to the appendage border to identify cases where the cutting plane identifies a shape containing the PVs. The ellipse fitting is achieved using the method described by Fitzgibbon et al. [48]. Given the vector form of the implicit second-order polynomial specification of the ellipse,

$$F_{\boldsymbol{a}}(\boldsymbol{z}) = \boldsymbol{z} \cdot \boldsymbol{a}, \tag{1.5}$$

where $\boldsymbol{z} = [x^2, xy, y^2, x, y, 1]$ and $\boldsymbol{a} = [a, b, c, d, e, f]^T$ are the variables and parameters vector, respectively. Then the parameter vector $\boldsymbol{a}$ is the solution of an optimisation problem where the sum of squared algebraic distances for the point to the conic is minimised:

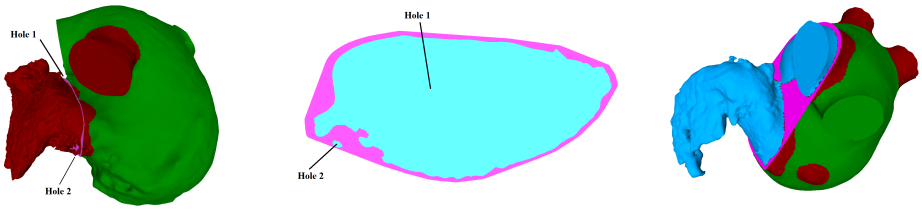
$$\min_{\boldsymbol{a}} \mathcal{D}_A(\boldsymbol{a}) = \min_{\boldsymbol{a}} \sum F_{\boldsymbol{a}}(\boldsymbol{z})^2. \tag{1.6}$$

The goodness of fit is evaluated as:

$$G = \frac{1}{N} \sum_{i=1}^{N} (F_{\boldsymbol{a}}(\boldsymbol{z_i}))^2 = \frac{\mathcal{D}_A(\boldsymbol{a})}{N}. \tag{1.7}$$

Empirical analysis suggests that cases with $G < 200$ should exclude the cutting plane that identifies a shape containing the PVs. Hence, if a case with $G \geq 200$ happens (Figure 1.7c), the method returns to the skeletonisation step and repeats the entire procedure.

Once the appendage is successfully cut, the LAA skeleton is detected using the convex hull of the appendage. The skeleton of the detected appendage is illustrated in Figure 1.5h.



(a) (b) (c)

Figure 1.7: (a) and (b) A cut along a trabecula generates multiple holes. (c) The cutting plane identifies a shape containing the PVs with a ellipse goodness of fit $G = 3,700$.

1.2.3 Phase 3: LAA Feature Extraction

Upon successful identification of the appendage, our method proceeds to automatically compute several features by analysing both the shape and skeleton of the LAA. The extracted features are summarised in Table 1.1.

Ostium

The LAA ostium, also LAA orifice, stands for the opening through which the appendage communicates with the LA. It is quantitatively described using the perimeter, area and eccentricity. To compute the ostium perimeter, denoted as L_O, the algorithm sums the length of the edges of the LAA that belong to the border identified by the cutting plane. The LAA shape is closed along the border identified by the cutting plane, forming a closed region. The ostium area, denoted as S_O, is computed by summing the area of the enclosed faces. The ostium eccentricity, denoted as e, is determined by fitting an ellipse to the border vertices, similar to the method used in the fourth step of Phase 2. Figure 1.8a and Figure 1.8b illustrate the ostium and the corresponding fitted ellipse. Additionally, the method computes the ratio between S_O and the area of the appendage, denoted as S_{LAA}. This ratio, denoted as $\mathcal{S}$, provides an informative measure of the ostium's size relative to the overall appendage surface.

Volume

The LAA volume, V_{LAA}, is computed by closing the LAA shape along the border, forming a solid region. Given the natural variability in heart sizes among individuals, the LAA volume is normalised by dividing it by the total

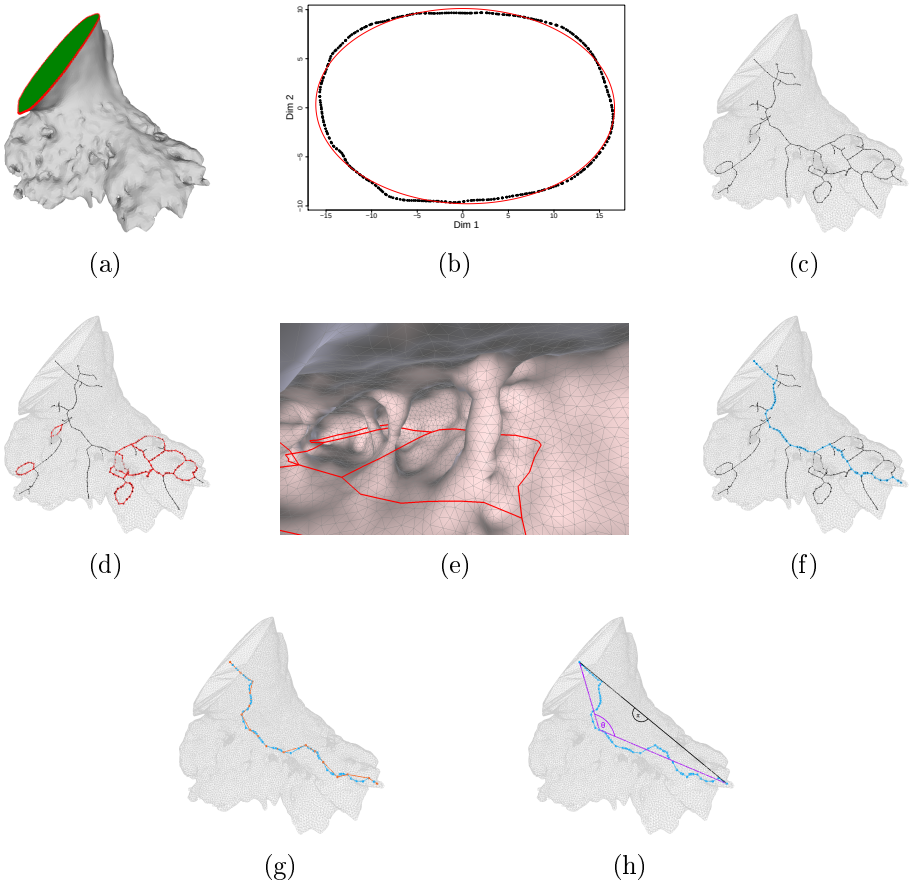


Figure 1.8: (a) The red path is used to compute the length of the ostium perimeter. The green faces are used to compute the area of the ostium. (b) The ellipse fitted to compute the eccentricity of the ostium. (c) The skeleton of the appendage. (d) The red path highlights the trabeculae (e) by showing how skeletal loops help to identify them. (f) The blue path highlight the main path. (g) The orange path shows the shortest path that preserves at least 95% of the length of the main path. (h) The length of the blue and black paths (arch-cord) are used to compute the tortuosity, while θ and π are used to compute the curvature.

shape volume, representing this ratio as $\mathcal{V}$. Normalizing the LAA volume allows for meaningful comparison of the LAA size across different heart shapes, thus enabling a more comprehensive analysis of the anatomical variations.

Trabeculae

A trabecula is a beam that connects opposite surfaces of the appendage anatomically represented by the pectinate muscle. Since the method works on data concerning the interior of the LAA surface, specifically the blood volume, it can use the skeleton structure to compute the number of trabeculae, denoted as N_T. A 3D graph analysis of the skeleton can detect the number of loops in its structure, which serves as a proxy for the number of trabeculae. A loop envelopes a trabecula Figure 1.8e. For the sake of simplicity, we will use the terms trabecula and loop interchangeably.
N_T is initially set to zero, and the algorithm performs a 2-core decomposition [159] of the skeleton. This process involves iteratively removing the vertices with a degree equal to one. The 2-core decomposition eliminates the skeleton queues while retaining both the trabeculae and any bridge between them. The analysis focuses on the skeleton nodes with a degree higher than two, as these nodes are more likely to belong to at least one loop. For each node with a degree higher than two, we select the shorter path, consisting of a minimum of three edges, that links the node with itself. These identified paths represent the loops connected to each other. We increment N_T by the number of paths and remove them from the skeleton. The remaining nodes belong to isolated loops and bridges. To remove the nodes belonging to the bridges (which are no longer linked to any loop), the algorithm repeats the 2-core decomposition, retaining only the nodes of the isolated loops. For one of the remaining vertices, the algorithm computes the shorter path linking the node to itself, thereby identifying the loop. N_T is increased by one, and the points in the identified loop are removed. The last step is repeated until no nodes remain in the skeleton. The trabeculae structure is illustrated in Figure 1.8d.

Main Path

The main path is identified as the longest possible route, measured in terms of Euclidean distance, among all the various shorter paths that connect the initial vertex located at the start of the LAA skeleton (specifically, at the LAA ostium) to each of the queue vertices, i.e. the terminal vertices of the skeleton. In other words, it is the path that covers the greatest distance when compared to all other available routes connecting these specific points. The main path is depicted with a blue line in Figure 1.8f.
Having obtained the length of the main path, denoted as L_{MP}, and considering

the straight line from the proximal end of the skeleton to the tip of the main path 1.8h, with a length of L'_{MP}, the tortuosity of the main path, denoted as τ, is determined using the arc-chord ratio:

$$\tau = 1 - e^{\left(1 - \frac{L_{MP}}{L'_{MP}}\right)}. \tag{1.8}$$

A tortuosity of 0 indicates that the main path is a straight line, while values closer to 1 indicate greater curvature and torsion along the path. Additionally, to account for situations where several appendages might have a similar tortuosity of the main path but differ in the angle of their evolution, the curvature, denoted as κ, is computed as follow

$$\kappa = \left| \frac{\min_{M_i} \widehat{SM_iE}}{\pi} - 1 \right|, \tag{1.9}$$

where S and E represent the starting and ending points of the main path, respectively, and M_i is the i-th point of the skeleton lying between S and E. The angle composed by the triple (S, M_i, E) cannot exceed π; hence, the curvature is obtained as:

$$\kappa = 1 - \frac{\min_{M_i} \widehat{SM_iE}}{\pi}. \tag{1.10}$$

The smaller the angle $\widehat{SM_iE}$, the higher the curvature of the main path; thus, κ approaches 1 when the angle tends to 0, while $\kappa = 0$ indicates that the main path is a straight line, i.e. $\min_{M_i} \widehat{SM_iE} = \pi$. The angle $\min_{M_i} \widehat{SM_iE}$ is denoted with θ in Figure 1.8h.

Lastly, the method computes the absolute density of the main path, denoted as γ_{abs}, which represents the number of vertices of the main path that preserves at least the 95% of L_{MP} (Figure 1.8g). To facilitate comparisons between models with different lengths of the main path, the method also calculates the normalised density, denoted as γ_{rel}, obtained by dividing the absolute density by the total number of vertices on the main path.

1.3 Experimental Validation

The dataset used in this work consists of eighty-six ($n = 86$) 3D anatomical models from adult patients who underwent a CT scan at University College

London Hospital (London, UK). The patients were scanned for moderate coronary diseases but no known atrial pathology; hence, they were considered control subjects. All patients gave written informed consent. To ensure patient privacy, the data were anonymised before use. Our method successfully detected all appendages without the need for human intervention. The detected LAAs are shown in Table A.1 of Appendix A, together with their cutting plane that separates them from the corresponding LA.

Table 1.1 provides a summary of the features extracted by our method and Figure 1.9 shows a graphical representation. The mean volume of the detected appendages was approximately 6.6 cm^3, while the mean values for ostium perimeter and area were 7.1 cm and 3.6 cm^2, respectively. The number of trabeculae shows high variability among the appendages, with a mean of approximately 9 and a standard deviation of 7. Similarly, the complexity of the main path exhibits significant variation, with an average of approximately 19 nodes and a standard deviation of 6, where the edge linking two nodes is 5 mm circa.

Table 1.1: Extracted features of the LAA.

Feature	Symbol	Unit	Domain	Mean	Median	St.Dev.
Ostium Perimeter	L_O	cm	$\mathbb{R}^+$	7.108	6.965	0.909
Ostium Area	S_O	cm^2	$\mathbb{R}^+$	3.579	3.427	0.970
Ostium Area to LAA Area Ratio	$\mathcal{S}$		$(0,1)$	0.119	0.120	0.024
Ostium Eccentricity	e		$[0,1]$	0.757	0.775	0.076
LAA Volume	V_{LAA}	cm^3	$\mathbb{R}^+$	6.597	6.192	2.505
LAA Volume to Total Volume Ratio	$\mathcal{V}$		$(0,1)$	0.069	0.067	0.020
Number of Trabeculae	N_T		$\mathbb{N}_0$	8.674	7.000	7.218
Length of the Main Path	L_M	cm	$\mathbb{R}^+$	5.332	5.238	1.125
Tortuosity of the Main Path	τ		$[0,1)$	0.499	0.458	0.181
Curvature of the Main Path	κ		$[0,1)$	0.541	0.526	0.150
Absolute Complexity of the Main Path	γ_{abs}		$\mathbb{N}$	19.337	19.000	6.366
Relative Complexity of the Main Path	γ_{rel}		$(0,1]$	0.289	0.277	0.098

1.4 Discussions

We presented a comprehensive method to improve critical challenges in the morphological analysis of the LAA. These challenges encompass the subjective nature of identifying the boundary between the LAA and the LA, and

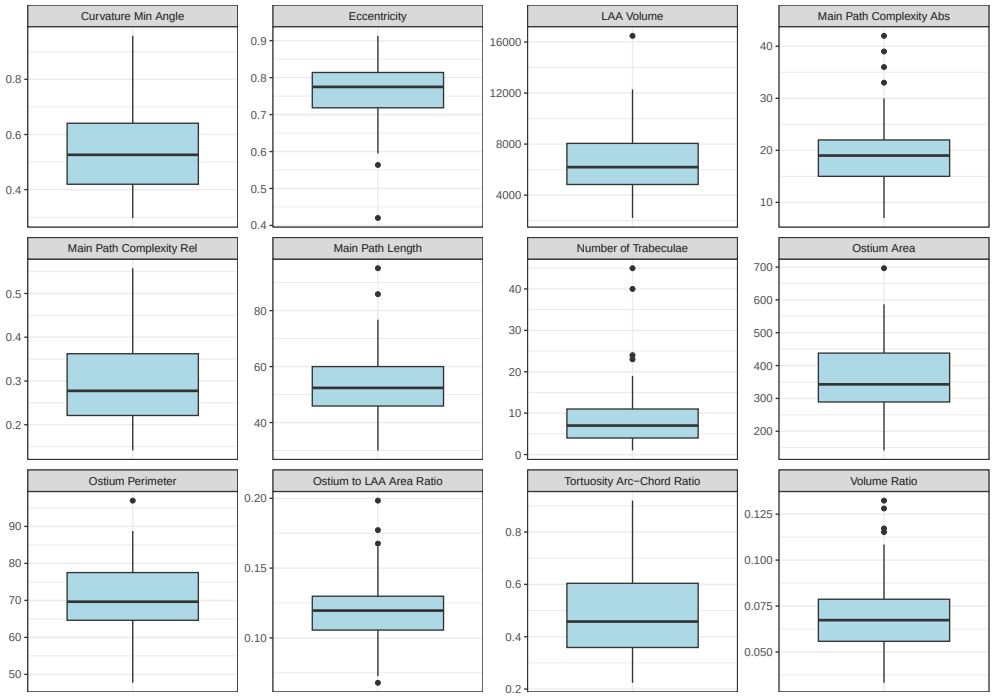


Figure 1.9: Box-plot of the features extracted by our method. Unit of measurement and domain of each feature is reported in Table 1.1.

the characterisation of LAA features. To effectively address these issues, our method is organised into four main phases. This chapter decribes the first three phases, each (except for the pre-processing phase) purposefully designed to tackle a distinct challenge. By focusing on these critical aspects, our approach ensures a more objective and accurate detection and characterisation of the LAA.

The LAA detection phase addresses the challenge of uncertainty in identifying the appendage from the LA. This phase consists of four steps, enabling automatic and robust LAA detection. In the first step, the system performs skeletonisation of the input shape to derive the skeleton. In the second step the system computes the distances between the shape and the skeleton. In step three, the method splits the shape into two groups based on a case-specific threshold, leading to smaller and larger distance groups. The affinity evaluation of faces within these groups refines and smooths the separation border. Finally, in step four, the method detects the auricula by analysing the group with smaller distances. It retains the section with the most complex structure of the inward skeleton and utilises PCA and ellipse fitting to identify the plane

(i.e. the LAA ostium) for cutting the appendage from the left atrium, ensuring consistency with clinical imaging techniques.
The deep learning methodologies [72, 140, 29] necessitate a training procedure for appendage detection, thus demanding the provision of ground truth data to the method, such as manual data labelling, a process susceptible to variations in user expertise and interpretation. Furthermore, achieving reproducibility with such methodologies poses challenges without access to the training datasets. Independently of the procedure employed to generate the 3D shape, our approach obviates the necessity for user input in LAA detection. This is accomplished through a plane identification process aimed at discerning the anatomical ostium of the appendage, guided by skeletal distances. Our method for LAA detection is entirely automated, distinguishing it from existing literature alternatives, which typically require some form of user input, such as specifying the LAA position [97] or utilising ground truth labelled data [72, 140, 29]. Furthermore, our approach has demonstrated efficacy in analysing complex LAA morphologies, accommodating geometries with trabeculae and non-smooth surfaces, and simplex LAA morphologies with no trabeculae and a not-so-evident junction from LA to LAA. These capabilities enhance accuracy and clinical relevance by incorporating a plane for segmenting the appendage, aligning with established clinical procedures.
Among the literature models capable of separating the appendage from the left atrium, Table 1.2 highlights those that additionally characterize the LAA with a minimum of six quantitative features, crucial for grouping appendages with similar morphological characteristics together. The table reveals two noteworthy findings: our approach stands out as the sole method to compute the eccentricity of the ostium; both Masci et al. [97] and our model are distinguished as the only two approaches to compute all features through automated procedures.

The LAA feature extraction phase aims at addressing the subjectivity issue in feature computation by adopting quantitative and reproducible methods. While features related to the appendage size, such as volume and ostium, are computed during the identification phase, those concerning the morphological complexity, such as the tortuosity of the main path and the number of trabeculae, are computed after the appendage is identified through an unsupervised network analysis of the skeleton. In contrast to the state-of-the-art practices, which often involve subjective identification of the number of lobes,

Table 1.2: Features extracted by the main literature methods are denoted by the 💻 symbol if their computation is automated, and by ✋ if they are manually derived. Features that are not extracted at all remain unmarked.

Feature	Our Method	Ernst et al. [45]	Budge et al. [20]	Wang et al. [152]	Yamamoto et al. [161]	Li et al. [88]	Masci et al. [97]	Rajiah et al. [119]
Ostium Eccentricity	💻							
Ostium Min. Diameter		✋	✋	✋	✋	✋		✋
Ostium Max. Diameter		✋	✋	✋	✋	✋		✋
Ostium Perimeter	💻			💻		✋	💻	✋
Ostium Area	💻			💻	✋	✋	💻	✋
Ostium Plane	💻							✋
Landing Zone				✋				
LAA area							💻	
LAA Volume	💻	✋	💻	✋	✋	💻	💻	
LA + LAA Volumne	💻							
Depth/Outline		✋	✋	✋	✋	✋		✋
Width			✋					
Main Path/Course	💻	✋		✋			💻	
Number Of Trabeculae	💻							
Number of Lobes/Branches		✋	✋	✋	✋	✋		
Number of twigs		✋						
Number of fine structures		✋						
Tortuosity	💻						💻	
Ostium to Landing zone distance				✋				
Curvature/Bend angle	💻			✋		✋		

branches and twigs affected by the imaging plane [45, 148, 132, 20], we adopt a different approach. Rather than characterising the auriculae based on the number of lobes, we utilised the extracted skeleton to describe the inner main path, enabling a more objective evaluation of its curvature. The procedure for evaluating the main path of the appendage and computing features of its complexity is explicitly described in our method. For quantitatively evaluating crenellation, we employed a network analysis of the skeleton to compute the number of trabeculae. An advantage over previous works [20, 66, 161, 88], our method to compute LAA features is fully unsupervised and relies on quantitative measures. We provide a detailed explanation of the procedure, resulting in a reproducible definition and computation for future studies.

While the proposed framework represents a significant advancement in LAA detection and its characterisation, there remain areas for further improvements. One crucial aspect is the need for large-scale validation and direct comparison between control and AF patient data. However, the feasibility of these enhancements relies on clinical data availability and access to extensive patient datasets.

1.5 Conclusions

The system described in this chapter addresses two of the three gaps in the morphological analysis of the left atrial appendage. Specifically, the method overcomes the uncertainty in identifying the auricula from the left atrium by

automatically detecting a cutting plane. It also uses quantitative and reproducible procedures to compute LAA features, addressing the subjectivity in the state-of-the-art methods to characterise the LAA.

We verified the effectiveness of the proposed method by applying it to a sample of 86 control patients. The tool detected all the appendages in the sample without human intervention.

In conclusion, the method represents a promising approach to achieve comparability in future studies of the morphological analysis of the left atrial appendage by allowing clinicians and researchers to detect and characterise LAA without human intervention, or rather, without subjective judgements.

Chapter 2

Advancement in Left Atrial Appendage Group Identity: A Feature-Based Clustering Approach

In this chapter, we present an overview of the current state of the art in classifying the left atrial appendage (LAA) and various machine learning algorithms aimed at classification and clustering.
Selecting the most suitable treatment option for atrial fibrillation (AF) patients depends partly on the morphological group to which the LAA is assigned. The prevailing classification method relies heavily on expert and clinician experience, which introduces subjectivity. There are different approaches to classifying appendix morphology. Moreover, the identification of the riskiest morphology varies across studies, even within those employing the same classification system. This highlights the need to develop a computerised LAA group identity assessment method that eliminates human intervention.
This chapter focuses on developing a feature-based clustering to assess the classification of LAA. We employed various clustering algorithms with many techniques. We present multiple partitions of the analysed sample and assess them based on a reliability measure obtained by applying several supervised classification methods. The outcome of our methodology is a data-driven assessment of the LAA group identity that distinguishes itself from state-of-the-art classifications based solely on shape similarities. It effectively groups patients with similar LAA features, enabling a more nuanced and precise classification.

This advancement aims to uncover more profound patterns and associations within the complexity of the LAA, contributing to a more comprehensive and objective assessment of the LAA group identity and its clinical implications.

Abstract

The traditional LAA classification relies heavily on expert opinion, which can be subjective. Here, we propose a feature-based clustering to group LAAs, employing various clustering algorithms and evaluating the partitions using supervised classification methods. This data-driven method provides a more precise classification based on multiple LAA features, revealing deeper patterns and associations within LAA morphology. Key findings include the identification of optimal partitions using K-means and Fuzzy K-means, which distinguished two LAA groups based on size and internal complexity. The study concludes by highlighting the need for future research to expand sample sizes and explore the clinical implications of these new LAA groups, particularly regarding clot-formation risk and treatment strategies.

2.1 Left Atrial Appendage Morphological Classification

Researchers, clinicians, and experts in the analysis of the LAA have endeavoured to identify the most high-risk morphologies. In their pursuit, they traditionally assigned appendages to groups based solely on their shape. While this approach is quick and convenient for classification, it overlooks crucial features of the appendage that are undoubtedly pertinent to therapy selection.
A more comprehensive understanding of LAA morphology is essential for optimizing therapeutic decisions. This entails considering the shape and factors such as size, location, and structural characteristics. By incorporating these additional features into the evaluation process, clinicians can make more informed choices regarding treatment strategies, ultimately improving patient outcomes. Therefore, researchers and practitioners must continue to refine their methods for assessing LAA morphology to ensure the most effective and tailored therapies are provided to those in need.

2.1.1 State-of-the-art LAA Classification

The classification of LAA morphology has seen several developments over the years. Wang et al. [152] introduced the widely adopted classification that categorised LAA shape into four groups: Chicken Wing, Windsock, Cauliflower, and Cactus. Particularly, a bend in the dominant lobe characterises a Chicken Wing LAA, which may also present secondary lobes or twigs. The Windsock LAA has a dominant lobe and varies across the location and the number of secondary and tertiary lobes. A Cauliflower LAA is short, has complex internal characteristics, and varies in ostium shape. The Cactus LAA has a dominant central lobe with secondary lobes extending from it in both superior and inferior directions. Their classification, which became the reference in the literature after Di Biase et al. [41], also classify the ostium by its shape into five groups: oval, triangular, round, foot-like and water drop-like.

Kimura et al. [78] refined this classification by incorporating the length and angle of the main lobe: LAA with a main lobe shorter than 4 cm is classified as Cactus if it has more than two lobes over 1 cm and as Cauliflower if none; while, LAAs with a main lobe longer than 4 cm are classified as Chicken Wing if it bend less than 100 degrees, as Windsock otherwise. Although they aimed for greater objectivity and reproducibility, their length and angle computation reliance on imaginary lines still introduced subjectivity based on the imaging plane.

Lacomis et al. [84] proposed a classification of three LAA types based on the orientation and location of the LAA tip. Two types differ in orientation - superiorly or inferiorly - but are always parallel to the main pulmonary artery and left cardiac border. The third type is oriented superiorly and turns medially between the main pulmonary artery and the left atrial body. Koplay et al. [82] expanded on this system by introducing a more detailed classification with five types and additional subtypes.

López-Mínguez et al. [93] categorised the LAA based on the ostium type and location. The first and second types are anterior and superior to the left superior pulmonary vein (LSPV), with the second type being closer to the LSPV orifice. The first type has a wide and flattened left lateral ridge (LLR), while the second type features a longer, narrower, and pointed LLR. The least common third type has the ostium located inferior and anterior to the left inferior pulmonary vein (LIPV) orifice, with an LLR similar to the second type. The limbus extending from the LAA ostium to the LIPV orifice is shorter in the

third and longer in the second types.
Some authors have classified the LAAs according to the number of lobes. Su et al. [137] classified the LAA as finger-like and stump-like, according to its intricate structure in bends, branches and twigs. Kaminski et al. [74] classified the LAA into four types, looking at the dominant lobe position, dimension and course and those of the secondary lobes. Finally, Wang et al. [150] classified the morphology into three groups based on the number of lobes: one, two or three.
Li et al. [88] proposed an alternative classification with seven types and six subtypes of LAA: Horseshoe, Cockscomb, Hand Finger, Paper Fan, Wing, Mountain, Hook, Wedge, Swan, Fork, and Ring.
Cressa et al. [34] proposed a step-by-step classification focusing on landing zone diameter and depth comparison first, then on the number of lobes, and finally on the dominant/useful lobe.
More recently, Słodowska et al. [138] aimed to simplify the classification by reducing the number of morphology types to three (Cauliflower, Chicken Wing, and Arrowhead). However, these criteria, like the previous ones, still require subjective judgement, leading to potential bias and subjectivity in classification.
The association between the risk of clot formation and LAA morphology, based on the classification of Wang et al. [152], has been extensively studied since Di Biase et al. [41]. Wu et al. [158], Yaghi et al. [160] reviewed these studies and found a lack of homogeneity in the results. Inconsistencies were observed where some morphologies showed no association with stroke risk and the group for which the difference in stroke risk is significant varied across different studies. This suggests that the reference classification might not be suitable for effectively discriminating the association between the LAA morphology and the thromboembolic risk.

2.1.2 Machine Learning Classification and Clustering Techniques

Machine learning, a subset of artificial intelligence, empowers computers to learn from data and make predictions or decisions based on that knowledge. This field has revolutionized various industries, from healthcare and finance

to self-driving cars and natural language processing. Machine learning algorithms can be broadly categorized into two main types: supervised learning and unsupervised learning, each with distinct characteristics and applications. [127, 14, 25].

The choice of a machine learning algorithm is contingent upon the nature of the available data, whether labelled or unlabelled. For example, the differentiation between classification and clustering tasks may appear subtle to individuals without specialized knowledge. These tasks encompass distinct objectives aligned with the data type under consideration. In a classification task, the user possesses *a priori* knowledge regarding the predefined groups to which each statistical unit is assigned. This type of data is termed *labelled data*. The user leverages this labelled information to instruct the machine learning algorithm. Essentially, the algorithm utilizes this knowledge to learn and discern patterns within the data. Subsequently, when presented with new, unlabeled data, the algorithm can accurately assign it to the group exhibiting analogous patterns guided by the acquired knowledge. Conversely, the clustering task entails systematically arranging statistical units into discrete groups. In this scenario, the user does not possess any foreknowledge concerning the correct assignment of individual units. This type of data is considered *unlabelled*. The fundamental objective of clustering is to identify inherent similarities among the data points and group them into collections based on these similarities. Importantly, there is no *a priori* guidance on which unit should belong to which group, making this task distinct from classification.

Supervised Learning Supervised learning involves training a machine on a dataset with the desired output or target variable. The model objective is to learn a mapping from input data to the correct output. Key characteristics of supervised learning include:

- *Labelled data*: the training dataset consists of input data and their corresponding output labels.
- *Objective*: the primary goal of supervised learning is to generalize from the training data and make accurate predictions on new, unseen data. The model performance is typically evaluated based on its ability to predict the output labels for new examples correctly.
- *Applications*: include image classification, speech recognition, sentiment

analysis, and regression tasks.

Several main techniques and algorithms are used in supervised learning, depending on the nature of the problem. Here are some of the primary techniques and algorithms in classification tasks:

- **Artificial Neural Networks** (ANN): draw inspiration from the intricate structure of the human brain [100]. These algorithms are potent data mining tools, particularly adept at pattern detection and predictive modelling, even when datasets are complex and incomplete. Recent applications of ANN have underscored their value in addressing real-world challenges, including the critical domain of fraud detection [5]. The architectural underpinning of ANN consists of three essential layers: the input, hidden, and output. This layered structure enables ANNs to capture intricate patterns within data, making them an asset for various applications. ANN stands out for its capability to uncover fundamental relationships within data, often yielding higher predictive accuracy than competing methodologies or human experts in specific use cases [22]. It is important to acknowledge the limitations of ANN. One of its principal drawbacks lies in its lack of transparency. ANN models often behave like enigmatic *black boxes*, making it difficult to explain the rationale behind their predictions. Even in cases with a seemingly simple architecture featuring only a single hidden layer, understanding why a specific data point is assigned to one class rather than another remains a formidable challenge [91].

- **K-Nearest Neighbours** (KNN): it operates on the foundational premise that similar data points are proximate to each other, a principle encapsulated in the adage that *like things are close to other like things* [49]. KNN functions by assessing the distances between a specific data sample and all the samples within the dataset. Subsequently, it identifies and selects the K closest samples, where K represents the specified number of neighbours. In classification tasks, the class label most frequently appearing among these selected neighbours is designated the prevailing class for the queried data sample. KNN is easy to use and has an expeditious calculation time. Furthermore, a noteworthy attribute of this algorithm is its nonparametric nature, signifying that it refrains from imposing assumptions about the underlying data distribution. This quality renders KNN

adaptable to a variety of datasets. However, it is essential to acknowledge that the efficacy of KNN is contingent upon the quality of the data. It necessitates well-structured and informative data to yield meaningful results. One of the primary challenges associated with KNN is determining an optimal value for K, which is the number of neighbours to consider. The choice of K can substantially impact the algorithm performance. Another limitation of KNN stands in applications where data points are positioned near classification boundaries, rendering them equivocal. In such instances, KNN may exhibit suboptimal classification accuracy, as the algorithm struggles to definitively assign these data points to one specific category or another.

- **Naïve Bayes** (NB): it is a probability-based classification algorithm based on Bayes' theorem [7, 81]. NB operates within a Bayesian framework, making use of conditional probabilities. It assigns probabilities for each of the classes. In simpler terms, it answers the question: *What's the probability that a data point belongs to a specific class, given the features we've observed?*. One of the fundamental assumptions of NB is that the features are mutually independent conditional on the class. This assumption is known as conditional independence. While this assumption is *naïve* and rarely holds in the real world, it simplifies the modelling process significantly. NB classifiers find significant application in text classification tasks, where documents or pieces of text are categorized into different classes or categories. It is particularly effective in tasks such as spam detection [125]. During the training phase, the algorithm learns from labelled data, calculating the probability of each class and the conditional probability of each feature given to each class. Hence, the algorithm transitions to the prediction phase, where it evaluates the probability of each class given the observed features using Bayes' theorem, ultimately selecting the class with the highest probability as the output. One of the notable strengths of Naïve Bayes is its rapid training and convergence time. It can quickly adapt to new data and is suitable for real-time applications. However, its *naïve* assumption of feature independence is a significant weakness, which can lead to suboptimal or inaccurate classifications, particularly when strong dependencies exist between features.

- **Random Forest** (RF): it derives its name from its composition, which encompasses a multitude of individual decision trees, functioning as an ensemble [68, 19]. Within this algorithm, each tree within the Random Forest contributes a class prediction, and the class that accumulates the majority of votes among these trees is designated as the model-predicted class. The foundational concept and motivation underlying the RF classifier are rooted in the notion of the *wisdom of crowds*. This concept is predicated on the belief that a substantial assembly of relatively uncorrelated models, represented by the constituent decision trees, collectively outperforms the predictive capacity of any individual tree. It posits that amalgamating a collective of weak learners engenders a strong learner. It is imperative to acknowledge that the RF classifier, while delivering robust performance, is characterized by greater complexity, making it relatively intricate to interpret.

- **Support Vector Machines** (SVM): introduced by Boser et al. [17] as a training algorithm for optimal margin classifiers. SVMs aim to find hyperplanes that can separate different groups or classes, which can be linear or nonlinear. SVMs seek to maximize the distance between the closest data points and the hyperplane, and this line is called the decision surface. In a 2D space, the decision surface separates classes, with margins between them. The data points closest to these boundaries are known as support vectors and play a crucial role in constructing the SVM classifier. Initially developed as a linear algorithm, SVMs have been extended to handle nonlinearly separable classes through techniques like the kernel trick [155]. Strengths of SVMs include good out-of-sample generalization when appropriate training parameters are selected. The introduction of the kernel trick adds flexibility, allowing SVMs to work with datasets that do not follow a linear pattern. However, SVMs also have some limitations. They lack transparency in describing the contribution of individual data features to the hyperplane, primarily due to high dimensions. Choosing the right kernel and parameter settings can be challenging, as their effectiveness varies for different problems. SVMs are also sensitive to irrelevant features in the dataset, which can lead to incorrect or suboptimal classification results.

- **Multinomial Logistic Regression** (MLR): it is often called multino-

mial logit. It embraces the principles of logistic regression, where the relationship between the independent variables and the categorical outcome is modeled through the logit function [83]. However, the categorical outcome variable can have three or more levels in the multinomial setting. The core of multinomial logistic regression lies in its ability to estimate the probability of each category relative to a reference category, considering the effects of the independent variables. This estimation is achieved by constructing separate logistic regression models for each category, with one category as the reference against which the others are compared. Fitting a multinomial logit model involves estimating parameters that describe the relationship between the independent variables and the log odds of each category relative to the reference category. These parameters are then used to predict each category probability for new observations based on their predictor values. Multinomial logistic regression finds applications across diverse fields, ranging from social sciences and economics to healthcare and marketing. Despite its utility, multinomial logistic regression poses certain challenges and considerations. The interpretation of model coefficients becomes more nuanced in the multinomial setting, requiring careful examination of the odds ratios and their associated confidence intervals.

Once the classification of new data is performed, evaluating the quality of the classification algorithm is crucial to assess its performance and suitability for a specific task. Several metrics and techniques can be used to gauge the effectiveness of a classification algorithm. Many of them are based on a tabular representation of the results, called **Confusion Matrix**. A confusion matrix shows the number of true positives, true negatives, false positives, and false negatives, providing a comprehensive view of the classifier performance [136]. Here are some of the most employed metrics:

- **Accuracy**: Accuracy is a straightforward metric that calculates the ratio of correctly predicted instances to the total number of instances. While easy to interpret, it may not be suitable for imbalanced datasets. [108]
- **Precision**: Precision, also known as positive predictive value, measures the proportion of true positive predictions out of all positive predictions. It is a valuable metric when false positives are costly. [58]
- **Recall** (Sensitivity): Recall, or true positive rate, calculates the ratio of

true positive predictions to the actual positive instances. It is important when false negatives are a concern. [58]

- **F1 Score**: The F1 score is the harmonic mean of precision and recall. It provides a balanced assessment of a classifier performance, especially when there is an imbalance between classes. [58]

- **Area Under the Receiver Operating Characteristic** (ROC-AUC): ROC-AUC is a metric used to evaluate the classifier ability to distinguish between positive and negative classes. A higher ROC-AUC score indicates better performance. [62]

- **Precision-Recall Curve**: The Precision-Recall curve illustrates the trade-off between precision and recall for different threshold values, providing insights into the classifier performance in imbalanced datasets. [38]

- **F-beta Score**: The F-beta score is a generalization of the F1 score that gives more weight to either precision or recall, depending on the beta value. Higher beta values favor recall, while lower values favour precision. [147]

- **Logarithmic Loss** (Log Loss): Logarithmic loss measures the uncertainty of a classifier predictions. Lower log loss values indicate more confident and accurate predictions. [36]

- **Matthews Correlation Coefficient** (MCC): MCC measures the quality of binary classifications that account for both true and false positives and negatives. It ranges from -1 (perfect disagreement) to +1 (perfect agreement). [98]

The choice of evaluation metric depends on the specific characteristics of the classification problem, including the class distribution, the importance of different types of errors, and the goals of the classification task. Using a combination of these metrics is common to get a more comprehensive assessment of a classification algorithm performance.

Unsupervised Learning Unsupervised learning involves training a model on unlabelled data without predefined output labels. Instead of learning from

explicit examples, the model task is to uncover hidden patterns or structures within the data. Key characteristics of unsupervised learning include:

- *Unlabelled data*: the training dataset comprises only input data.
- *Objective*: the primary goal of unsupervised learning is to reveal hidden structure, group similar data points, or reduce the complexity of the data for further analysis. It is often used for explanatory data analysis.
- *Applications*: unsupervised learning is employed in tasks such as clustering, reducing the dimensionality of data, and anomaly detection.

As for supervised learning, several main techniques and algorithms are used, depending on the nature of the problem. Here are some of the primary techniques and algorithms in clustering tasks:

- **Exclusive Clustering**: In this approach, data points are assigned exclusively to a single cluster, and membership in more than one cluster is prohibited. An example is the K-means algorithm [94, 64, 92]. This method offers a straightforward and efficient way to classify a given dataset into a predetermined number of clusters, denoted as k. The core concept of K-means involves establishing k cluster centres, one for each cluster. The placement of these centroids plays a crucial role in the success of the algorithm, as different initial centroid positions can yield varying outcomes. Hence, an optimal strategy is to position these centroids as far apart as possible. The algorithm initializes the k centroids, distributing them across the dataset. Each data point in the dataset is then associated with the nearest centroid. This step continues until all data points have been assigned to a cluster, constituting an initial grouping. With the initial grouping, the algorithm recalculates k new centroids as each cluster barycenter (center of mass) formed in the previous step. Data points are reassigned to the nearest centroid, and new centroids are recalculated. Recalculating new centroids and reassigned data points continues until the centroids no longer change, indicating that they have stabilized.
  K-means does not guarantee that it will find the optimal configuration in terms of the global optimum. It finds a local optimum based on the initial placement of centroids and the specific dataset. Indeed, the outcome of the K-means algorithm can vary depending on the initial placement

of centroids. If the initial centroids are chosen poorly or the dataset has inherent complexities, the algorithm may converge to a suboptimal solution. To mitigate this, K-means is often run multiple times with different initializations, and the best result is selected by minimizing the sum of squared distances from data points to their assigned centroids. This approach is known as the **K-means++** initialization [4].

- **Overlapping Clustering**: Overlapping clustering allows data points to belong to multiple clusters with varying degrees of membership. Fuzzy K-means is an example of this approach [43, 12]. It is an extension of the traditional K-means clustering algorithm, allowing for a more nuanced assignment of data points to clusters. In traditional K-means, each data point is assigned exclusively to the nearest cluster, resulting in a hard assignment. On the other hand, Fuzzy K-means introduces the concept of *fuzziness*, where each data point has a probability of belonging to each cluster. This allows data points to be somewhat between clusters or exhibit ambiguity in their membership. The probability of a data point belonging to a cluster could be a function of the distance. For example, the probability might be inversely proportional to the distance from the cluster centroid. Fuzzy K-means uses weighted centroids for each cluster. These weighted centroids consider the probabilistic assignments of data points. Data points with higher probabilities of belonging to a cluster contribute more to the cluster centroid calculation. The initialization, iteration, and termination processes in Fuzzy K-means are like traditional K-means. It typically starts with initial centroids, iteratively refines these centroids, and terminates when convergence criteria is reached. It is worth noting that traditional K-means can be seen as a special case of Fuzzy K-means where the probability function is simplified to be 1 if the data point is closest to a centroid and 0 otherwise, resulting in a hard assignment. Fuzzy K-means is valuable when dealing with data that exhibits a degree of ambiguity or when a soft clustering assignment is more appropriate than a strict, exclusive assignment. It has applications in areas where data points may have multiple cluster memberships with varying degrees of confidence.

- **Hierarchical Clustering**: This approach arranges data points into a dendrogram, a tree-like structure, where data points with greater simi-

larity are grouped together within clusters. There are two main types of hierarchical clustering: agglomerative [65] and divisive [76]. The agglomerative hierarchical clustering commences by designating each data point as an individual cluster (if there are N data points, there will be N initial clusters, each containing a single data point). The cluster distances (or similarities) are initialised based on the pairwise distances between the individual data points. The algorithm identifies the two clusters that exhibit the closest proximity, determined by a predefined distance metric or similarity measure. The choice of distance metric is contingent on the data characteristics and the specific problem under consideration. These two clusters are merged into a new cluster, resulting in one less cluster in the hierarchy. After merging two clusters, the algorithm updates the distances (or similarities) between the newly formed cluster and the remaining clusters is essential. The procedure continues this iterative process by consistently merging the closest clusters and updating distances until all data points are organised into a solitary cluster, culminating in creating a hierarchical structure. Divisive hierarchical clustering works oppositely: it starts with all data points in a single cluster and computes pairwise distances; then selects the cluster with the largest internal dissimilarity and divides it into two clusters; finally, it recalculates the distances between the new clusters and the existing clusters. The procedure is repeated until each data point is in its cluster. Recomputing the distances after an aggregation or a cluster division requires applying linkage methods [134, 153], which dictate how distances are computed between the new cluster and existing ones. Common linkage methods encompass *single linkage*, where the distance between clusters is defined as the minimum distance between any data point in one cluster and any data point in the other; *complete linkage*, where the distance is defined as the maximum distance between any data points in the two clusters; *average linkage*, where the distance is determined as the average of all pairwise distances between data points in the two clusters. The outcome of hierarchical clustering is a dendrogram, which visually represents the hierarchical relationships between clusters. It empowers users to choose the number of clusters at varying granularity levels, catering to specific analytical requirements.

- **Probabilistic Clustering**: Probabilistic clustering employs a fully prob-

abilistic approach to delineate clusters. Mixture models are examples of this approach. A mixture model is a probabilistic framework for data representation, wherein a dataset is interpreted as a composite assembly of distinct component distributions [120]. Each of these component distributions corresponds to an individual cluster within the data. The fundamental idea is that data points within each cluster originate from a shared underlying distribution, while the entire dataset represents a mixture of these distinct distributions. The selection of the appropriate distribution depends on the inherent characteristics of the data and the specific problem under consideration. Mixture models offer well-established techniques for statistical inference. This implies that one can effectively estimate model parameters, including those defining the component distributions, by harnessing methodologies such as the Expectation-Maximization (EM) algorithm [40]. Such statistical underpinnings offer robust and principled means of learning model parameters from the data. A strength of mixture models lies in their soft clustering feature; indeed, it refrains from rigidly binding data points to single clusters. Instead, each data point embodies a probability distribution across the available clusters, accommodating uncertainty in cluster assignments. This stands in stark contrast to hard clustering, which necessitates unequivocal allocations. The Gaussian Mixture Model (GMM) [120] emerges as a specific instance within the realm of mixture models. In a GMM, every component distribution takes the form of a Gaussian (normal) distribution. GMMs are central in cluster analysis and density estimation, underscored by their efficacy and approachability. They exhibit remarkable flexibility, with the ability to emulate an extensive array of data distributions. Their adaptability empowers them to capture clusters of varying shapes and sizes within the data. The applications of GMMs span diverse domains, encompassing image segmentation, speech recognition, and anomaly detection [15]. Moreover, they find utility in generative modelling, a context in which they can craft novel data samples that faithfully mimic the training data distribution.

2.1.3 Network Analysis and Community Detection

Network analysis is a robust and versatile methodology employed to explore and comprehend intricate system configurations, interconnections, and be-

haviours [154]. Its applicability extends across a wide spectrum of domains, encompassing social networks, biological systems, transportation networks, and the vast expanse of the World Wide Web. By embracing network analysis, we equip ourselves with a powerful lens capable of revealing concealed intricacies, associations, and insights that elude conventional data analysis techniques.

Network analysis entails meticulously examining networks comprising nodes (also known as vertices) and edges (or links) that establish connections between these nodes. Nodes serve as the elemental entities within the network, representing individual components. Meanwhile, edges encapsulate the relationships or interactions between these entities. This approach to modelling real-world systems as networks yields profound insights into their intricate dynamics, operational attributes, and the emergence of unique properties.

The foundations of network analysis are firmly grounded in graph theory, a mathematical framework meticulously tailored to model and dissect networks [13]. Graphs can be of distinct types, including directed or undirected, weighted or unweighted, and may exhibit an array of topological characteristics. Among these features are the small-world attributes, the prevalence of scale-free degree distributions, and the revelation of underlying community structures.

Network analysis introduces fundamental metrics, particularly centrality measures, instrumental in identifying pivotal nodes within a network. Prominent centrality metrics encompass degree centrality, betweenness centrality, and eigenvector centrality. These metrics unveil the nodes influencing information dissemination, communication pathways, or network dynamics.

Additionally, network analysis proves indispensable in elucidating the mechanisms governing information dissemination, behavioural diffusion, or the propagation of diseases across interconnected networks. Various models, such as the epidemic spreading and information diffusion models, offer the means to simulate and scrutinize these intricate processes.

Notably, many real-world networks manifest a scale-free degree distribution, where a minute fraction of nodes presents high degrees while the majority exhibit lower degrees. A profound understanding of the structure of these scale-free networks plays an instrumental role in forecasting their resilience and vulnerability under various perturbations.

Network analysis extends its utility beyond its conventional applications and can be harnessed for data clustering. Though it may not adhere to the con-

ventional paradigms of clustering techniques such as K-means or hierarchical clustering, it offers a unique approach. Community detection algorithms, employed within network analysis, orchestrate the clustering of nodes into communities or clusters based on their interconnectivity [56, 31, 87]. These communities delineate subsets of nodes that exhibit dense intra-connectivity, providing profound insights into the inherent modular structure of networks.

2.2 LAA Data-Driven Group Identification

In this section, we comprehensively examine the methodologies employed for the optimal data-driven clustering of LAA morphologies. First, we use some unsupervised clustering methodologies like K-means and Fuzzy K-means. The crux of this approach hinges upon the necessity of prespecifying the requisite number of clusters into which the LAA entities shall be partitioned. This necessitates empirically determining the optimal cluster count through diverse techniques, including applying silhouette and elbow methods. Another facet of our methodological arsenal is the partitioning of LAAs by deploying hierarchical clustering with agglomerative and divisive methodologies. An inherent attribute of this technique is its innate agnosticism towards the pre-specification of cluster numbers, contingent upon the imposition of specific conditions related to dendrogram cut and distance matrices. Finally, we use several network representations of the LAAs to detect communities based on different detection methodologies.

As the number of partitions of the sample data equals the number of methodologies applied, we keep the partitions where all clusters/communities have at least two members and the number of groups varies from two to six. Then, we apply several supervised machine learning algorithms. These algorithms serve the dual function of affirming the robustness of the proposed classifications and identifying the most reliable. Finally, in-depth scrutiny is directed towards pairwise and entire comparisons on the kept partitions. The outcome of this rigorous comparative analysis culminates in identifying primary clusters and salient morphological attributes intrinsic to the cluster itself.

This comprehensive methodological framework forms the foundation of our systematic approach to characterising LAA identity groups. In the following experiment, we will consider a dataset with N data points, denoted as $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, where each $\mathbf{x}_n$ represents a data point in the feature

space, say $\mathbf{x}_n$ is a vector of f components, where each f-th component refers to one of the F features extracted in Chapter 1. Given the feature domains are different, we used the standardised representation of the feature to find the partitions, i.e. $\mathbf{Z} = \frac{\mathbf{X}-\mu_{\mathbf{X}}}{\sigma_{\mathbf{X}}}$. For the sake of simplicity, we will refer to the clusters/communities/groups independently by using the notation C, where C_k represents the k-th cluster/community/group of the partition found through a given method.

2.2.1 LAA Data-Driven Partitions

The aim of this subsection is to find a partition of the sample data with different clustering approaches. Table 2.1 shows the numerous clustering approaches with the differnt techniques and parameters used to find the partitions.

Table 2.1: Clustering approaches utilised to find LAAs partition.

Method	Techniques	Parameters	Number of Partitions
K-means	Elbow Silhouette Gap Statistic		3
Fuzzy K-means	Partition Coefficient Modified Partition Coefficient Partition Entropy Silhouette	fuzzy parameter (m) $\in \{1.5, 2.0, 2.5\}$	15
	Fuzzy Silhouette	$\alpha = 1$, fuzzy parameter (m) $\in \{1.5, 2.0, 2.5\}$	
Hierarchical Clustering	Agglomerative, Complete Linkage Agglomerative, Average Linkage Agglomerative, McQuitty Linkage Divisive, DIANA		4
Gaussian Mixture Model		Assumptions in $\Sigma_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^\top$ configuration	14
Community Detection in Positive Correlated Netwok	Edge Betweenness InfoMap Louvain Leiden Leading Eigen		5
Community Detection in Minimum Spanning Tree	Edge Betweenness InfoMap Louvain Leiden Leading Eigen	Single, Average, and Complete Linkage in MST definition	15
Community Detection in Planar Maximally Filtered Graph	Edge Betweenness InfoMap Louvain Leiden Leading Eigen		5
Positively Correlated Network	Edge Betweenness InfoMap Louvain Leiden Leading Eigen		5
Leave-one-Out Robustness Network	Edge Betweenness InfoMap Louvain Leiden Leading Eigen		5
K-fold CV Robustness Network	Edge Betweenness InfoMap Louvain Leiden Leading Eigen		5

K-means

K-means clustering is one of the most widely employed unsupervised machine learning algorithms used to partition a given dataset into a predetermined number of groups, denoted as k, as specified by the user. In the context of this work, we utilize the classical k-means algorithm originally introduced by Hartigan and Wong [64]. This method defines the total within-cluster variation as the sum of squared distances between the data points and their respective cluster centroids. Mathematically, it can be represented as follows:

$$W(C_k) = \sum_{z_i \in C_k} (z_i - \mu_k)^2$$

Here z_i represents the i-th observation belonging to the cluster C_k, and μ_k is the mean value of the data points assigned to cluster C_k. The objective is to assign each observation n to a cluster such that the sum of squared distances from the observation to its assigned cluster center μ_k is minimized, where the total within-cluster sum of squares, WSS, is calculated by summing $WSS(C_k)$ for all clusters, i.e.

$$\arg\min_{K} \sum_{k=1}^{K} WSS(C_k). \tag{2.1}$$

This metric serves as an indicator of the compactness or quality of the clustering.

The initial step in utilizing k-means clustering involves specifying the number of clusters (k) that will be formed in the final result. Determining the optimal value of k can be challenging. However, several methods can help identify this parameter:

- **Elbow Method**: This technique suggests that the optimal number of clusters is where the total within-cluster sum of squares (WSS) is minimized. It calculates the WSS for various values of k and selects the one with the smallest WSS. To identify the optimal number of group, $\hat{k}$, the user can plot the WSS values against the number of clusters and look for a bend point, signifying the appropriate number of clusters. We choose $\hat{k}$ such as the smallest k such that $0.95 \geq \frac{WSS(k)}{WSS(k-1)} \geq 1.00$

- **Silhouette Method**: The silhouette method [123] assesses how well each data point fits within its cluster and quantifies the quality of the clustering. It computes the average silhouette of observations for different

k as follows:

$$S_n = \frac{b_n - a_n}{\max\{a_n, b_n\}}, \tag{2.2}$$

where a_n (b_n) is the average distance between unit n and the other units in the same (nearest) cluster. The optimal k, $\hat{k}$, is the one that maximizes the average silhouette over a range of possible values for k.

- **Gap Statistic**: The gap statistic [141] compares the total intra-cluster variation for different k values derived from a null reference distribution, which represents a distribution with no discernible clustering. The reference distribution is generated using Monte Carlo simulations of the sampling process. This process entails calculating the range for each variable z_n within the dataset and generating values for m data points uniformly distributed within the interval defined by the minimum and maximum values. For both the observed and the reference data, the total intra-cluster variation is computed across different k values. The gap statistic for a given k is defined by the equation:

$$GAP_m(K) = E_m^*\{log(W_k)\} - log(W_k), \tag{2.3}$$

where E_m^* is the expectation under a sample size m from the reference distribution and is determined through bootstrapping B copies of the reference datasets. The gap statistic quantifies the deviation of the observed W_k value from its expected value under the null hypothesis. The estimate of the optimal clusters, $\hat{k}$, is the smallest k such that $GAP(k) \geq GAP(k+1) - s_{k+1}$.

Fuzzy K-Means

Fuzzy k-means is an extension of the traditional k-means clustering algorithm that allows data points to belong to multiple clusters simultaneously to varying degrees.

Like k-means, the algorithm starts by randomly initialising k cluster centroids or using a predefined method. Each data point is assigned to every cluster, with a membership value representing the degree of belongingness of each data point to each cluster. These membership values are often randomly assigned or initialised.

After the initial assignment of membership values, the centroids of the clusters

are updated based on the membership values through a weighted mean, where each data point contributes to the centroid based on its membership value:

$$\mu_{C_k} = \frac{\sum_x \omega_k(z)^m \cdot x}{\sum_z \omega_k(z)^m}; \tag{2.4}$$

hence, data points with higher membership values influence the centroid's position more. m is the hyper-parameter that controls how fuzzy the cluster will be. The higher it is, the fuzzier the cluster will be.
Once the centroids are updated, the membership values for each data point are recalculated based on the new centroid positions. This step adjusts the degree of belongingness of each data point to each cluster based on the updated centroids. The procedure continues until convergence, i.e., when the centroids no longer change significantly between iterations or when a predetermined number of iterations is reached.
Once convergence is reached, the algorithm outputs the final clustering, where each data point has membership values indicating its degree of belongingness to each cluster.
Fuzzy k-means, like classical k-means, aims to minimise a predefined objective function, which typically includes the sum of squared distances between data points and centroids, weighted by the membership values. This objective function measures the overall fit of the clusters to the data while considering the fuzzy nature of the memberships. Reducing fuzziness to exclusive group belonging is reached by assigning each data point to the cluster for which the data point presents the maximum membership value. As K-means, fuzzy K-means also needs the number of clusters to be specified. Several criteria are usually used for this scope. Apart from the **Silhouette** method, already described in the K-means, other techniques include:

- **Partition coefficient**: it measures the degree of fuzziness or overlap between clusters [11]. It is typically computed using a formula that considers the membership values assigned to each data point and the distance between data points and centroids. It is computed as follows:

$$PC = \frac{1}{N} \sum_k \sum_n v_{nk}^2 \tag{2.5}$$

  where v_{nk} represents the membership value of unit n in cluster k. A higher partition coefficient indicates better separation between clusters

and a clearer delineation of cluster boundaries. Therefore, once the partition coefficient is computed for every k, $\hat{k}$ is identified as the k with the greater partition coefficient.

- **Modified partition coefficient**: it is a modification of the partition coefficient, where the denominator is adjusted to account for the fact that each data point should have a single dominant cluster membership in an ideal clustering scenario [37]. The formula to compute it is:

$$MPC = 1 - \frac{k}{k-1}(1 - PC) \tag{2.6}$$

This modification helps penalise situations where data points have high membership values across multiple clusters, leading to a more balanced assessment of cluster quality.

- **Partition entropy**: it measures the degree of fuzziness or uncertainty in assigning data points to clusters by calculating the membership values entropy across clusters [10]. Higher entropy values indicate greater uncertainty in cluster assignments, implying more fuzziness or overlap between clusters. Partition Entropy is computed as:

$$PE = -\frac{1}{n}\sum_{n}\sum_{k} v_{nk} \cdot ln(v_{nk}), \tag{2.7}$$

where v_{nk} represents the membership value of unit n in cluster k. The optimal cluster number is the k that minimises PE.

- **Fuzzy Silhouette**: it is an extension of the silhouette coefficient previously described. The fuzzy version considers the distance between a data point and its cluster centroid (intra-cluster distance) and between the data point and the nearest neighbouring cluster centroid (inter-cluster distance), considering the fuzzy memberships [24]. Fuzzy Silhouette is computed as follows:

$$FS = \frac{\sum_n (\mu_{v_{1,n}} - \mu_{v_{2,n}})^\alpha S_n}{\sum_n (\mu_{v_{1,n}} - \mu_{v_{2,n}})^\alpha} \tag{2.8}$$

where S_n is the silhouette of unit n, and $\mu_{v_{1,n}}$ ($\mu_{v_{2,n}}$) are the first (second) largest element of the n-th column of the fuzzy partition matrix. α is a weighting coefficient which regulates the relative importance of data objects in overlapping area. This coefficient ranges from -1 to 1, where a value closer to 1 indicates better separation between clusters and higher cohesion within clusters.

Hierarchical Clustering

Hierarchical clustering organizes data points into a hierarchical structure of clusters, where similar data points are grouped at lower levels of the hierarchy, and more dissimilar clusters are grouped at higher levels.

Agglomerative The steps of an agglomerative hierarchical clustering algorithm are as follows:

1. organise the units into N clusters, each containing a single data point. Let $C_i = \{z_i\}$ be the initial cluster for each data point z_i.
2. compute the distance matrix D between each pair of clusters. Let $d(C_i, C_j)$ represent the dissimilarity between clusters C_i and C_j.
3. identify the cluster pair with the smallest dissimilarity, $(C_p, C_q) = arg\min_{i \neq j} d(C_i, C_j)$, and merge them into a new cluster $C_{pq} = C_p \cup C_q$.
4. update the dissimilarity matrix by removing the no longer present cluster C_p and C_q and adding the newest C_{pq} such as $d(C_{pq}, C_k) = g\left(d(C_p, C_k), d(C_q, C_k)\right)$, where $g(\cdot)$ is a rule defined by the chosen linkage criterion, such as single (minimum distance), complete (maximum distance), average (average distance), ...
5. repeat steps 3. and 4. until only one cluster remains. This results in a hierarchy that a dendrogram can represent.

In this work, we use three different linkage criteria: complete linkage, which defines the distance between two clusters as the minimum distance between all the units in one cluster and the ones in the other cluster; average linkage, which defines the distance as the average of the distances between the units in the two clusters; McQuitty linkage, which is like the average linkage where the distances are weighted considering the sizes of the clusters.

Divisive The procedural steps of a divisive hierarchical clustering algorithm are:

1. organise the units into a single cluster. Let $C_0 = \{z_n\}$ with $n = 1, 2, \ldots, N$ denote the initial cluster.
2. compute the distance matrix D between each pair of clusters. Let $d(C_i, C_j)$ represent the dissimilarity between clusters C_i and C_j.

3. identify the cluster with the highest intra-cluster dissimilarity. This cluster is considered the most heterogeneous and is selected for splitting.

4. divide the selected cluster into two smaller clusters based on the chosen splitting criterion.

5. repeat steps 3. and 4. until the desired number of clusters is reached or all data points form individual clusters.

By setting the distance threshold $\hat{d}$, the dendrogram can be cut to a specific height, providing the final clusters. We work on a distance matrix built from a correlation matrix, such as

$$d(i,j) = \sqrt{2 \cdot (1 - \rho(i,j))}. \tag{2.9}$$

Distance defined this way assumes value 2 when two observations i and j are perfectly negatively correlated ($\rho(i,j) = -1$), while it assumes value 0 when they are perfectly positively correlated ($\rho(i,j) = 1$). It assumes value $\sqrt{(2)}$ when there is a null correlation between the two units. Therefore, a good cut criterion when we want to keep clusters that share similar characteristics is to cut the dendrogram in correspondence of a distance that avoids negatively correlated units from being clustered together; hence, we use $d(i,j) = \sqrt{(2)}$ to cut the dendrograms.

Gaussian Mixture Models

Gaussian Mixture Model (GMM) assumes that the data points are generated from a mixture of several Gaussian distributions. Each Gaussian component represents a cluster in the data, and the goal of GMM clustering is to estimate the parameters of these Gaussian distributions to fit the data best. Each data point has a probability of belonging to each cluster. This allows for soft clustering, where data points are not rigidly assigned to a single cluster but instead exhibit a degree of uncertainty in their cluster assignments.
Let's assume a GMM with the following form

$$p(\mathbf{x}|\Psi) = \sum_k \omega_k g(\mathbf{x}|\mu_k, \Sigma_k), \tag{2.10}$$

where ω_k are the mixture weights, $g(\mathbf{x}|\mu_k, \Sigma_k)$ are the Gaussian components, with their mean μ_k and covariance matrix Σ_k, and Ψ expresses the collection of all component parameters, i.e. $\Psi = (\omega_k, \mu_k, \Sigma_k)$.

Table 2.2: Parameterisations of the within-group covariance matrix $\mathbf{\Sigma}_k$ and the corresponding geometric characteristics. Source Scrucca et al. [129].

Model	$\mathbf{\Sigma}_k$	**Distribution**	**Volume**	**Shape**	**Orientation**
EII	$\lambda\mathbb{I}$	Spherical	Equal	Equal	
VII	$\lambda_k\mathbb{I}$	Spherical	Equal	Equal	
EEI	$\lambda\mathbf{A}$	Diagonal	Equal	Equal	Coordinate axes
VEI	$\lambda_k\mathbf{A}$	Diagonal	Variable	Equal	Coordinate axes
EVI	$\lambda\mathbf{A}_k$	Diagonal	Equal	Variable	Coordinate axes
VVI	$\lambda_k\mathbf{A}_k$	Diagonal	Variable	Variable	Coordinate axes
EEE	$\lambda\mathbf{D}\mathbf{A}\mathbf{D}^\top$	Ellipsoidal	Equal	Equal	Equal
EVE	$\lambda\mathbf{D}\mathbf{A}_k\mathbf{D}^\top$	Ellipsoidal	Equal	Variable	Equal
VEE	$\lambda_k\mathbf{D}\mathbf{A}\mathbf{D}^\top$	Ellipsoidal	Variable	Equal	Equal
VVE	$\lambda_k\mathbf{D}\mathbf{A}_k\mathbf{D}^\top$	Ellipsoidal	Variable	Variable	Equal
EEV	$\lambda\mathbf{D}k\mathbf{A}\mathbf{D}^\top k$	Ellipsoidal	Equal	Equal	Variable
VEV	$\lambda_k\mathbf{D}k\mathbf{A}\mathbf{D}^\top k$	Ellipsoidal	Variable	Equal	Variable
EVV	$\lambda\mathbf{D}k\mathbf{A}k\mathbf{D}_k^\top$	Ellipsoidal	Equal	Variable	Variable
VVV	$\lambda_k\mathbf{D}k\mathbf{A}\mathbf{D}^\top k$	Ellipsoidal	Variable	Variable	Variable

The fitting algorithm widely used to estimate the GMM is the Expectation-Maximization (EM) algorithm, which determines the optimal parameters of the GMM by iteratively updating them and maximise the likelihood $\mathbb{L}(\Psi) = \Pi_n g(\mathbf{x}_n, \Psi)$. In particular, after the parameter initialisation Ψ^i, the procedure evaluates the conditional expectation for each Gaussian component given the current parameters at every data sample (E-step). Hence, the parameter Ψ^{i+1} is updated for maximisation of likelihood to matching the observed $\mathbf{x}$ values. The covariances matrices can be parametrised as $\Sigma_k = \lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^\top$, where λ_k controls the k-th ellipsoid volume, $\mathbf{D}_k$ is orthogonal and controls the k-th ellipsoid orientation, and $\mathbf{A}_k$ is a diagonal matrix which specifies the shape of the density contours [6, 26]. According to Scrucca et al. [129], there are 14 possible models with different geometric characteristics, which can be obtained by constraining shape, orientation and volume to be equal or variable across groups. These models are summarised in Table 2.2.

After fitting all these models, the best model is selected using the Bayesian Information Criterion (BIC) as the twice approximation of the log Bayes factor

[50]. The formula of BIC is

$$BIC_M = 2\log_{\left(\mathbb{L}(x|\hat{\Psi})\right)} - m_M \log(n) \approx 2\log(p(x|M)) + c, \tag{2.11}$$

where $\log_{\left(\mathbb{L}(x|\hat{\Psi})\right)}$ is the maximised mixture log-likelihood for the model M, and m_M is the number of independent parameters to be estimated in the model. As BIC approximates the Bayes factor with maximum integrated likelihood ($p(x|M)$), the best model is the one with the largest BIC.

Network Analysis and Community Detection

In this section, we delineate the methodological framework employed for network analysis to identify clusters of LAAs. Specifically, within the context of network analysis, each LAA constitutes a node, denoting a fundamental unit of representation. Meanwhile, edges within the network paradigm articulate the relational structure between nodes, delineating the strength of association between respective LAAs. Let's denote the set of LAAs with $N = \{1, \ldots, n\}$ and edges as $e_{i,j}$, where i and j are nodes in N. We characterise the edges with a measure derived from pairwise correlation. The LAA network is undirected, meaning there is no direction from one node to another but rather a strength that connects two nodes. We use three representations of such a network: Complete Network (CN), Minimum Spanning Tree (MST), and Planar Maximally Filtered Graph (PMFG).

Complete Network A Complete Network (CN) represents the densest connectivity structure achievable within a given set of nodes. In this network composition, each node is connected to every other node within the network. Consequently, if the number of nodes equals N, then the total number of edges is given by $N\frac{N-1}{2}$.
A primary reduction of the Complete Network is achieved by considering only the edges with positive weights and utilising them as the weights for the network. This approach is justified by the objective of the analysis, which is to identify groups of LAAs that exhibit similar feature values, indicating positive correlations among them. By excluding edges with negative weights representing negative correlations, we retain only the edges relevant to our analysis. This reduced network will be referred to as CN^+.

Minimum Spanning Tree A minimum spanning tree (MST) is a subset of a dense network. It is the minimal representation of the network where all the

vertices are connected together without forming any cycles [117]. Hence, for a network composed of N network it will have $N-1$ edges.
An easy algorithm for finding the MST is as follow:

0. Initialise an empty MST.
1. Sort all the edges in ascending order of their weight (distance).
2. Select the edge with the smallest weight.
3. Check if the selected edge forms a cycle within the MST formed so far. If the cycle is not formed, include this edge. Else, discard it.
4. Check if there are $N-1$ edges within the MST. If not jump to [2.]; otherwise, STOP.

MST is closely associated with single linkage cluster analysis (SLCA) because all the information needed for SLCA of a set of points is encapsulated within their MST[59]. A modification of the MST algorithm involves selecting the next edge based on the minimum average distance, known as average linkage cluster analysis (ALCA) [146], rather than selecting the next edge based on the minimum edge weight. Similarly, it is feasible to select the next edge based on the maximum distance, known as complete linkage cluster analysis (CLCA).

Planar Maximally Filtered Graph Another special network is the planar graph, which can be drawn on a flat surface without its edges crossing over each other. The Planar Maximally Filtered Graph (PMFG) is a type of planar graph where the edges connecting the most similar elements are added first. [145] An easy algorithm for finding the PMFG is as follows:

0. Initialise an empty PMFG.
1. Sort all the edges in descending order of their weight (correlation).
2. Select the edge with the highest weight.
3. Check if the selected edge ensures the planarity of the PMFG. If the planarity condition holds, include this edge. Else, discard it.
4. Check if there are $3\,(N-2)$ edges within the PMFG. If not, jump to [2.]; otherwise, STOP.

The PMFG preserves the MST as a subgraph.

Robustness Metric Here, we use the CN^+ and a leave-one-out technique to obtain a robustness matrix, R^1, characterising the links. The $n \times n$ robustness matrix is initialised as a null matrix, $r^1_{ij} = 0 \; \forall i, j \in N$. We iteratively drop an appendage and detect the communities described in the following paragraph. For all appendages clustered together, we increase the corresponding cells of the robustness matrix by 1. At the end of the procedure (i.e. after the n-th iteration), the matrix is divided by $n-2$ (i.e. the total number of networks that could contain the pair) to normalise the robustness matrix in the range $[0, 1]$. The leave-one-out procedure results in a robustness matrix where 0 means that the two appendages never appear in the same group, while 1 means the pair is always present in the same class. Of course, an appendage is always present with itself, so the main diagonal of the robustness matrix is set to 1.
The same procedure is performed with k-fold cross-validation. We initialise the robustness matrix R^k as a null matrix. In this case, we randomly drop 20% of the units and perform the community detection on the remaining 80% of the units. We use a leave-one-out technique to detect the communities by keeping uncertainties in community detection and having a robust assignment. The remaining 20% units are assigned to the cluster that minimises the mean square error to its centroid. We increase the matrix entries by 1 if the corresponding units are clustered together. After repeating the procedure k times, the robustness matrix is divided by the total number of times two units can be clustered together, k.

Community detection The recognition of groups within networks is achieved through community detection, a field that has seen the development of several methods over the years. Main measures to evaluate the goodness of a division are:

- **Modularity**: it is defined as

$$M = \frac{1}{2m} \sum_{i,j} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j), \tag{2.12}$$

  where m is the number of edges, a_{ij} is the element of the adjacency matrix A for units i and j, k_i (k_j) is the degree of i (j). δ is a function that assumes value 1 if i and j belong to the same community, $c_i = c_j$, 0 otherwise. [31]

- **Edge betweenness**: it is defined as

$$EB = \sum_{i \neq j} \frac{g_{iej}}{g_{ij}}, \tag{2.13}$$

where g_{ij} is the total number of path between vertices i and j, while g_{iej} is the number of paths which pass through edge e. [51]

In this work, we utilized the following community detection techniques:

- **Edge Betweenness**: it quantifies the number of shortest paths that pass through an edge. The concept of edge-betweenness-based community structure detection rests on the observation that edges connecting distinct modules are likely to have high edge-betweenness scores, as all shortest paths from one module to another must traverse through them. By iteratively removing edges with the highest edge betweenness scores and recalculating the edge betweenness, it is possible to construct a hierarchical map known as a dendrogram, which forms a rooted tree representation of the graph. In this dendrogram, the leaf nodes correspond to individual vertices, while the root node represents the entire graph. [110]

- **Louvain**: it is based on the modularity measure and employs a hierarchical approach. Initially, each vertex is assigned to its community. In each iteration, vertices are reassigned to communities in a local, greedy manner: each vertex is moved to the community that maximizes its contribution to the overall modularity. When no further vertex can be reassigned, each community is treated as a single vertex, and the process repeats with the merged communities. The algorithm terminates when only one vertex remains or when no further increase in modularity can be achieved in a step. [16]

- **Leiden**: it is a hierarchical clustering algorithm that iteratively merges communities into single nodes by greedily optimizing modularity, and this process is repeated within the condensed graph. It is a modification of the Louvain algorithm designed to address situations where communities are not well-connected. The Leiden algorithm periodically breaks down communities into smaller, more well-connected ones. This adaptive refinement process helps improve the quality of community detection by ensuring that communities are more cohesive and accurately represent the underlying structure of the network. [142]

- **Leading Eigen**: it operates by computing the eigenvector corresponding to the largest positive eigenvalue of the modularity matrix. Subsequently, it partitions vertices into two communities based on the sign of the corresponding elements in the eigenvector to improve modularity. This process iteratively unfolds until there is no longer an improvement in modularity. [109]

- **InfoMap**: it utilizes random walks and compresses the conditional information of a random walk on the network into Huffman codes. InfoMap seeks to maximize the information of the random walk process by optimizing a partition function that minimizes the entropy of movement between and within communities. Infomap employs a deterministic greedy search algorithm to explore the space of possible partitions, which is further refined using a simulated annealing approach. [121, 122]

2.2.2 Clustering Validation

Once the data sample has been partitioned into groups using the previously described methodologies, we validate them using supervised methodologies. Given the numerous methods, techniques, and parameter sets employed in finding these partitions, we only consider divisions consisting of at least two but no more than six groups. Furthermore, additional restrictions are imposed regarding the minimum number of units in each group, ensuring that all clusters contain at least two units.
Selected partitions are validated through supervised classification methods. In this context, for each partition, i.e., the considered output of the unsupervised methods, we treat membership as the ground truth label, train the supervised classifiers, and finally compute the accuracy of the test set. Greater accuracy in such methods would suggest the goodness of the partition and, hence, of the corresponding clustering method [21, 99].

K-Nearest Neighbours The application of KNN implies the following steps:

1. The first step is to divide the data into a training and a test set, generally of proportion 80:20.

2. For each test unit, compute the distance between it and all the units in the training data. Euclidean distance is the most used.

3. Select K nearest neighbours.

4. Identify and assign the most repeated label among the K nearest neighbours.

The choice of K can impact the algorithm performance; hence, tuning it is a generally good practice.

Support Vector Machine The SVM model we use to classify the LAAs is trained using the C-classifier. A one-vs-one approach is employed to deal with a multi-class classification problem. The one-vs-one method trains multiple binary classifiers, each distinguishing between a pair of classes. During prediction, these binary classifiers vote on the class label, and the class with the most votes is assigned to the sample. The C parameter in SVM controls the trade-off between maximizing the margin and minimizing the classification error. A higher value of C means a smaller margin will be accepted if the decision function is better at classifying all training points correctly. In contrast, a lower C will encourage a larger margin and, therefore, a simpler decision function at the cost of training accuracy. In this context, we use the radial kernel $K\left(x, x'\right) = e^{-\gamma\left(x-x'\right)^2}$, where γ parameter defines how far the influence of a single training example reaches, with low values meaning far and high values meaning close. We employed a cross-validation technique to tune the C and γ hyperparameters and select the optimal values.

Multinomial Logistic Regression In multinomial logistic regression, the objective is to model the probability π_k to belong to a specific group k. After choosing a reference group, say $k = 1$, the logit of π_i, with $i \neq 1$ is computed as follows:

$$logit(\pi_i) = log\left(\frac{\pi_i}{\pi_1}\right) = \mathbf{X}^\top \beta i \tag{2.14}$$

After fitting, the estimated probability can be expressed as:

$$\pi_i = \begin{cases} \frac{1}{\sum_{k\neq 1}\left(1+\exp\left(\mathbf{X}^\top\beta_k\right)\right)} & \text{if } i = 1, \\ \frac{\exp\left(\mathbf{X}^\top\beta_k\right)}{1+\sum_{k\neq 1}\exp\left(\mathbf{X}^\top\beta_k\right)} & \text{if } i \neq 1 \end{cases} \tag{2.15}$$

New observations are assigned to the group for which the highest probability is computed.

Random Forest Random Forest algorithms combine the predictions of multiple decision trees trained on bootstrapped samples with feature bagging to create an ensemble model. A Random Forest algorithm consists of the following steps:

1. Set the three main hyperparameters. The node size, ns, determines the minimum node cardinality, i.e., the minimum number of units belonging to the leaf node; we use $ns = 1$. The number of trees, nt, specifies the number of decision trees to be included in the random forest ensemble; we set $nt = 1000$. The last parameter is the number of features sampled, nf, which determines the number of features randomly selected at each split in a decision tree (bagging procedure), usually set to $\sqrt{f}$ where f is the number of the features.

2. Each decision tree in the ensemble is built using a bootstrap sample from the training set. Bootstrap sampling involves randomly selecting samples from the training set with replacements. A feature bagging is applied to each decision tree, ensuring more diversity in the dataset and reducing correlation among decision trees.

3. Each decision tree makes its prediction, and the class with the most frequent occurrences among the trees is chosen as the final prediction.

Artificial Neural Network The neural network architecture used in this work has a simple structure. It is composed of one input layer, one hidden layer and one output layer. The number of neurons in the input layer equals the number of features in the training data. The number of groups of a given partition gives neurons in the output layers. The number of neurons in the hidden layer is the mean of the neurons in the input and output layers.
Once the architecture has been defined, we use the rectified linear unit activation function (ReLU) to calculate the output of the neurons. Essentially, each artificial neuron receives signals from connected neurons and sends a new signal to other connected neurons. The signal is the output of the activation function. The strength of the signal at each connection is determined by a weight, which adjusts during the learning process.
Weights are updated through the resilient propagation algorithm (RPROP), which adjusts the step size for weight updates based on the gradient sign.

Naive Bayes Classifier Let's assume the effect of a particular feature in a class is independent of other features. We can use the Bayes Theorem to use the so-called Naïve Bayes Classifier.
Bayes theorem provides a way of computing posterior probability from $P(k)$, $P(x)$, and $P(x|k)$ such that:

$$P(k|x) = \frac{P(x|k)P(k)}{P(x)}, \tag{2.16}$$

where $P(k|x)$ is the posterior probability of class k given the predictor x, $P(c)$ ($P(x)$) is the prior probability of the class k (the predictor x), and $P(x|c)$ is the likelihood which is the probability of the predictor given class. Given that the denominator does not depend on the class k and that it is observed, $P(k|k)$ is proportional to the numerator, which can be seen as $P(k)\Pi_k P(x|k)$ given the naïve conditional independence and using the chain rule. So,

$$P(k|x) \propto P(k)\Pi_k P(x|k). \tag{2.17}$$

The naive Bayes classifier combines the Bayes model with a decision rule. We use the maximum a posteriori decision rule (MAP) to assign units to classes. Hence, the classifier is defined as:

$$\hat{k}(x) = \arg\max_k \left\{P(k)\Pi_k P(x|k)\right\}. \tag{2.18}$$

Partition Reliability The capabilities of the supervised algorithm are evaluated by using the **Balanced Accuracy**, BA, computed as follows

$$BA = \frac{\sum_k \frac{TP_k}{n_k}}{K}, \tag{2.19}$$

where TP_k and n_k are the correct assignments and the total number of units in group k and K is the total number of classes.
This index converges to the classical accuracy one when the classes are balanced, i.e., they have the same number of units. BA gives better results when classes are unbalanced.

2.2.3 Partitions Comparison

This section shows the methodologies used to compare the several partitions found. The main measures to compare partitions are based on a co-membership confusion matrix and a frequency cross-tabulation of objects. Let's $\mathcal{P}$ and $\mathcal{Q}$ be the two partitions we are interested in comparing, which are composed of P and Q number of clusters, respectively.

- **Co-membership confusion matrix**: all possible unit pairs can be organised in a 2×2 matrix that represent the overlap of the two partitions $\mathcal{P}$ and $\mathcal{Q}$. In particular, we can define a co-membership confusion matrix, $M_{\mathcal{P},\mathcal{Q}}$, as:

$$M_{\mathcal{P},\mathcal{Q}} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}, \tag{2.20}$$

where m_{11} is the number of pairs present in both partitions, m_{22} is the number of pairs present either in $\mathcal{P}$ nor in $\mathcal{Q}$, and m_{12} (m_{21}) are exclusive pairs present in $\mathcal{P}$ ($\mathcal{Q}$) but in $\mathcal{Q}$ ($\mathcal{P}$).
The similarity of the two partitions is evaluated through the following measures:

  - **F1-score**: it is the harmonic mean of two asymmetric measures. Let's precision be defined as, $P = \frac{m_{11}}{m_{11}+m_{21}}$, and recall as, $R = \frac{m_{11}}{m_{11}+m_{12}}$, then $F1$ is defined as:

$$F1 = \frac{2m_{11}}{2m_{11} + m_{12} + m_{21}}. \tag{2.21}$$

    $F1$ assumes value 1 when the two partitions are the same (clusters may be labelled differently) and 0 when the two partitions do not share any pair.

  - **Jaccard Index**: it is computed as:

$$J = \frac{m_{11}}{m_{11} + m_{12} + m_{21}}, \tag{2.22}$$

    the numerator describes the intersection of the two partitions, which means the number of pairs present in both partitions and the denominator is the total number of pairs present. J approaches 1 as the two partitions overlap perfectly, it equals 0 when the two partitions share no pair.

- **Frequency cross-tabulation of objects**: the N units must belong to one or another group in both partitions. The idea is to arrange the units in a matrix representing how many units are in the corresponding group of each partition. Let's $FXTO_{\mathcal{P},\mathcal{Q}}$ be the frequency cross-tabulation of objects defined as:

$$FXTO_{\mathcal{P},\mathcal{Q}} = \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1Q} \\ n_{21} & n_{22} & \cdots & n_{2Q} \\ \vdots & \vdots & \ddots & \vdots \\ n_{P1} & n_{P2} & \cdots & n_{PQ} \end{bmatrix}, \tag{2.23}$$

where n_{pq} with $p \in P$ and $q \in Q$ represents the number of units present in both cluster p of partition $\mathcal{P}$ and in cluster q of partition $\mathcal{Q}$.
The widely used measures to evaluate the similarity of the partitions through the frequency cross-tabulation of objects are:

- **Mutual Information**: it is an entropic measure computed as follows:

$$MI = H(\mathcal{P}) + H(\mathcal{Q}) - H(\mathcal{PQ}), \tag{2.24}$$

  where $H(\mathcal{P})$ and $H(\mathcal{Q})$ are the entropy in partition $\mathcal{P}$ and $\mathcal{Q}$, respectively, and $H(\mathcal{PQ})$ is the joint entropy of $\mathcal{P}$ and $\mathcal{Q}$.

- **Normalised Mutual Information**: it is the normalisation of the MI. Several normalization have been proposed in literature; here, we use the formulation with max entropy:

$$NMI = \frac{MI}{\max\{H(\mathcal{P}), H(\mathcal{Q})\}} \tag{2.25}$$

- **Adjusted Mutual Information**: it is a further normalisation of MI while evaluating distance from random partitions. It is computed as:

$$AMI = \frac{MI - E(MI)}{\max\{H(\mathcal{P}), H(\mathcal{Q})\} - E(MI)}, \tag{2.26}$$

  where $E(MI)$ is the expected magnitude of MI when $\mathcal{P}$ and $\mathcal{Q}$ are two independent, random disposition of units into groups.

2.3 Experimental Results

The method featured in Chapter 1 of this thesis enables the computation of various LAA attributes. These attributes exhibit differences in scale and units of measurement due to their diverse nature. Furthermore, some of these attributes pertain to similar characteristics, leading to potentially high levels of correlation among them. To address these issues, we standardized the LAA features and conducted pairwise comparisons to assess their correlation. In cases where a pair of features, both describing similar characteristics of the LAA, exhibited a correlation coefficient exceeding 0.75, we systematically eliminated the feature with the highest mean correlation to reduce redundancy.
In Figure 2.1a, we present a correlation plot among all the extracted features. Table 2.3 lists the removed features in comparison with its kept counterpart.

In particular, we removed the ostium area as it shows a higher average correlation compared to the ostium perimeter. Additionally, we found that one feature among curvature and tortuosity is sufficient to represent the evolving complexity of the main path, and we choose to keep the curvature feature due to its minimum value of $\mu(\rho)$. Furthermore, the absolute and relative complexity of the main path were also found to be correlated, and we retained the relative complexity feature. In Figure 2.1b, we present the correlation plot after removing the highly correlated features.

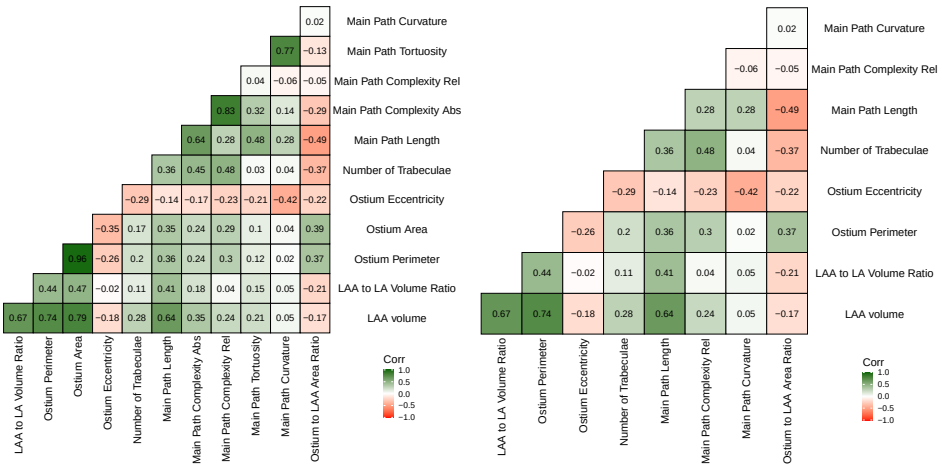


Figure 2.1: Correlation plot among features extracted by the method before (a) and after (b) the highly correlated features removal

Table 2.3: Removed variable by correlation evaluation

Removed Feature	Kept Feature	ρ	$\mu(\rho)_{\text{Removed}}$	$\mu(\rho)_{\text{Kept}}$
Ostium Area	Ostium Perimeter	0.96	0.32	0.30
Main Path Complexity Abs.	Main Path Complexity Rel.	0.83	0.31	0.19
Tortuosity Arc-Chord Ratio	Curvature Min. Angle	0.77	0.17	0.11

LAA Data-Driven Partitions

In this subsection, we show the results obtained by applying the methodologies described in the previous section to the same data used in Chapter 1, the 86 control patients (see Section 1.3 of this thesis for further details). Table 2.4 summarises the several approaches applied.

K-Means The results of the three strategies used to identify the optimal number of clusters in utilising the k-means algorithm are shown in Figure 2.2. The elbow method suggests that there are six distinct clusters within the LAA dataset (top), as six is the smallest number of groups that satisfy the condition $0.95 \geq \frac{WSS(k)}{WSS(k-1)} \geq 1.00$. In contrast, the silhouette method indicates the presence of only two clusters (bottom). However, it is worth noting that the GAP statistic fails to provide statistical support for rejecting the null hypothesis of data genetared by a single distribution, hence we have a single group (middle).

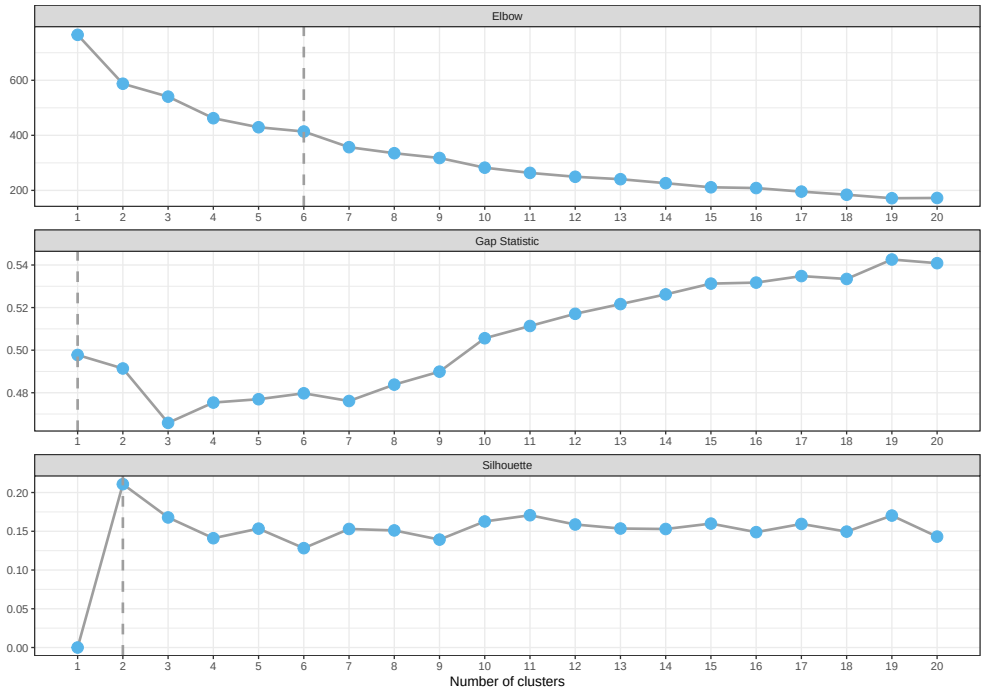


Figure 2.2: Number of clusters identified in K-means approach by elbow (top), gap statistic (middle) and silhouette (bottom) methods.

Fuzzy K-Means The results of the five strategies used to identify the optimal number of clusters in utilising the fuzzy k-means algorithm are shown in Figure 2.3. All methods suggest two distinct clusters within the LAA dataset but the modified partition coefficient with fuzzy parameter $m = 1.5$ indicates the presence of twenty clusters. In this case, one of the clusters has cardinality equal to 1, i.e. a LAA forms its own cluster. The rest of the methods identify two clusters of cardinality 46 and 40 (45 and 41) when fuzzy parameter is in

$\{1.5, 2.0\}$ ($\{2.5\}$). It may represent that a LAA changes its membership as fuzziness increases, showing an uncertainty in assigning it.

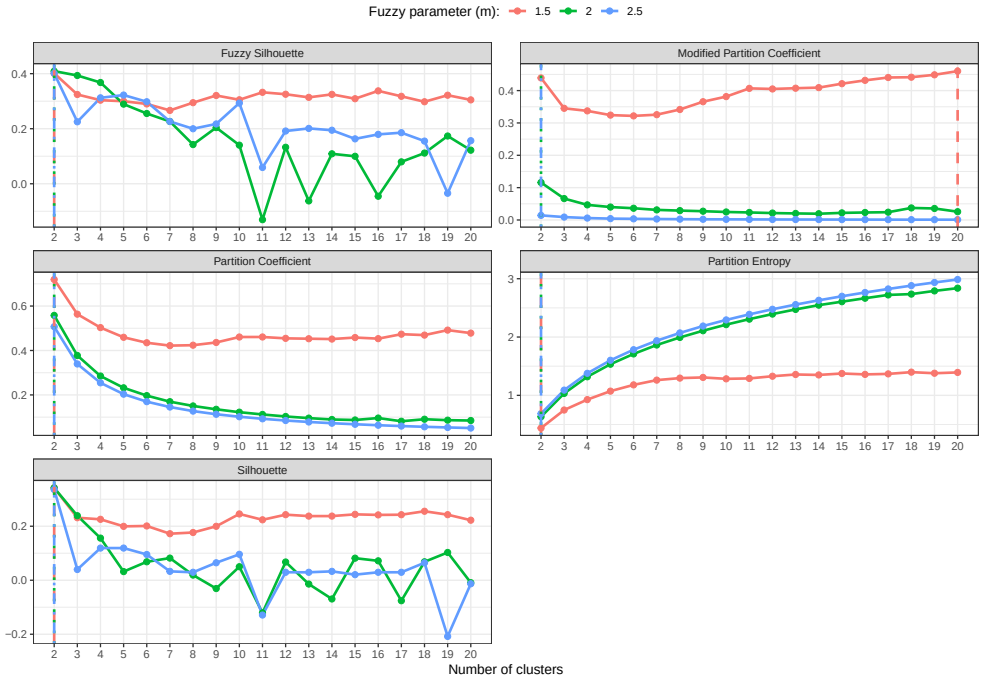
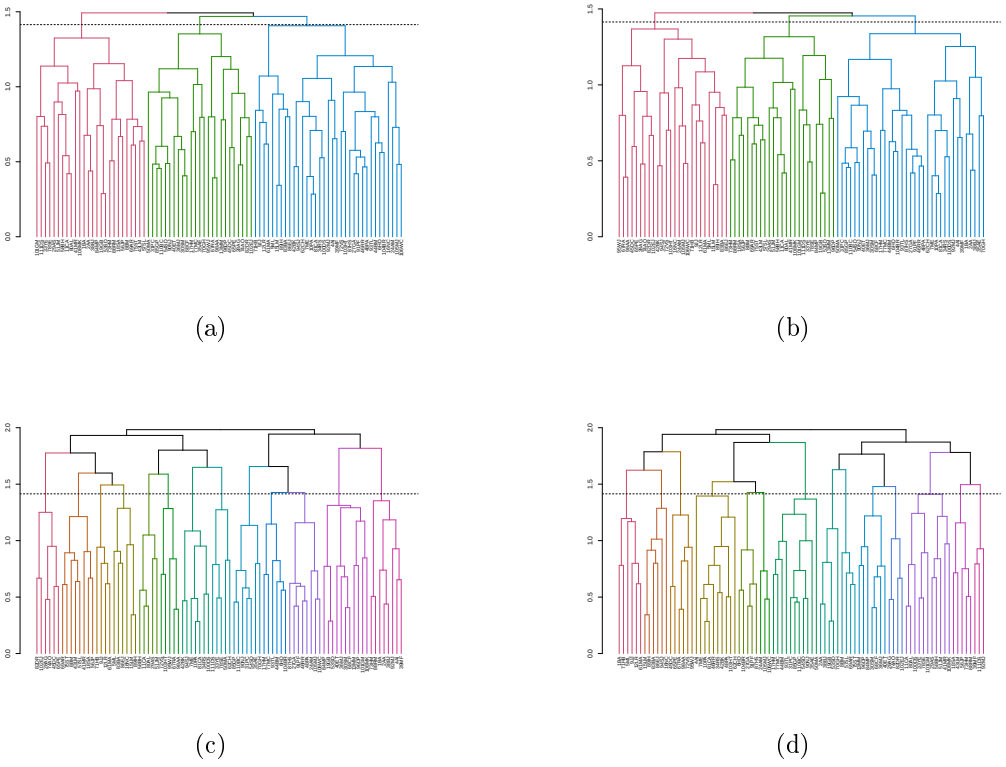


Figure 2.3: Number of clusters identified in Fuzzy K-means approach by Fuzzy Silhouette (top left), Modified Partition Coefficient (top right), Partition Coefficient (middle left), Partition Entropy (middle right), and Silhouette (bottom left). Colours shows results for three fuzzy parameters: $m = 1.5$ (red), $m = 2.0$ (green), and $m = 2.5$ (cyan)

Hierarchical Clustering The results of the four strategies used to identify the optimal number of clusters in utilising the hierarchical clustering are shown in Figure 2.4. Average (Figure 2.4a) and McQuitty (Figure 2.4b) linkages, both based on average distances, suggest three clusters. Complete linkage in agglomerative clustering (Figure 2.4c) and divisive hierarchical clustering (Figure 2.4d) indicate 13 and 14 groups, respectively.

Gaussian Mixture Models The results of the Gaussian Mixture Model classifier are shown in Figure 2.5. The three best GMM are VEE-2 (red point), VEE-3 (green point), and VVE-2 (blue point). The three models assume that the components have the same ortientation matrix and variable volume. The



(a) (b)

(c) (d)

Figure 2.4: Dendrograms of the hierarchical clustering approaches: (a) agglomerative with average linkage; (b) agglomerative with McQuitty linkage; (c) agglomerative with complete linkage; (d) divisive Diana. Horizontal lines represent the dendrogram cut at a $d = \sqrt{2}$, corresponding to a null correlation between units.

difference in the within-group covariance matrix parametrisation of the two models lies in the assumption of the components shape where VEE assumes equal shape while VVE does not.

Community Detection in Networks The results obtained by applying several community detection techniques in the five network configurations (CN, MST SLCA, MST ALCA, MST CLCA, PMFG) are briefly reported in Table 2.4. Among the 25 partitions, just three in the CN configuration and one in the PMFG configuration are composed of at least two and no more than six communities, each of them with a cardinality of at least two nodes. However, the modularity is null for the three divisions obtained in the complete configuration, thus suggesting that the community structure is not significantly

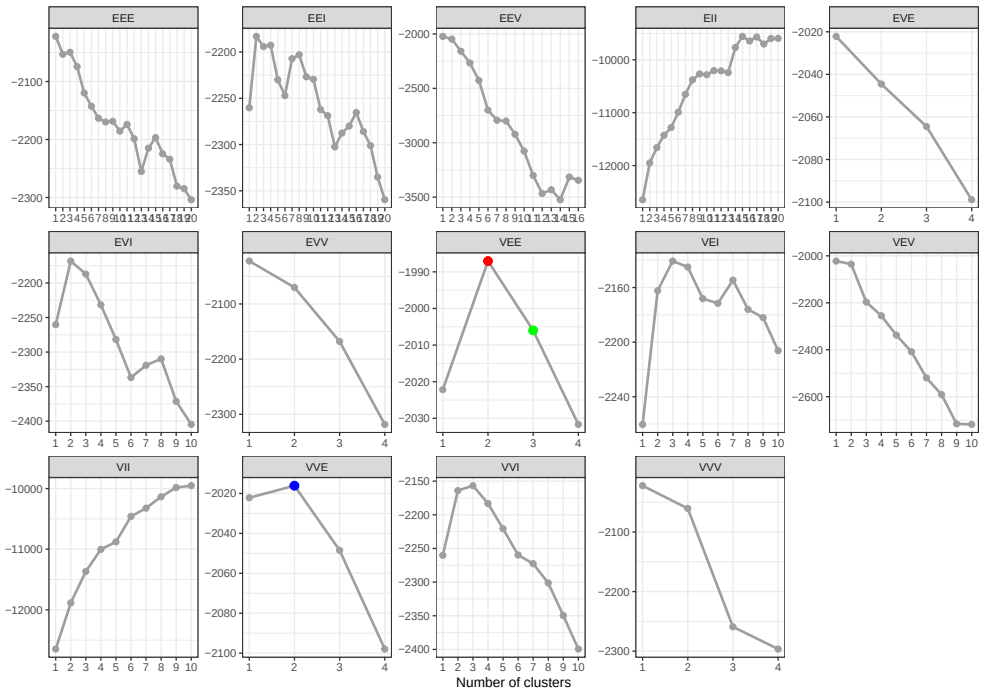


Figure 2.5: BIC values for combination of number of clusters and within-group covariance matrix parametrisation. Best three models are identified through red, green and blue points.

stronger than expected in a random network. On the contrary, the partition identified in the PMFG configuration with the Edge Betweenness method delineates clear and meaningful groupings of nodes, with relatively few connections between these groups. The modularity value is $M = 0.67$, suggesting that the observed division of the network into communities is much stronger than what would be expected in a random network. The PMFG configuration network with the Edge Betweenness-based communities is illustrated in Figure 2.6a.

Reducing the CN just by considering the positively correlated units, i.e. keeping the links with positive values, does not improve the CN results. Still, the leave-one-out procedure applied generally increases the modularity of the networks, while k-fold reduces it. Edge Betweenness and InfoMap cannot recognise communities. Louvain recognises four communities in the three configurations. Leiden identifies five communities in the positively correlated and k-fold configurations and 4 in the leave-one-out one. Finally, Leading Eigen detects three communities in the positively correlated and leave-one-out config-

urations and two in the k-fold configuration. Among the partitions composed of at least two and no more than six communities, each of them with a cardinality of at least two nodes, the highest modularity values are found in the leave-one-out configurations, with a peak for the Louvain method ($M = 0.40$), followed by the Leading Eigen ($M = 0.27$). These two networks and corresponding communities are shown in Figure 2.6b and Figure 2.6c.

Clustering Validation Twenty-four partitions among the previously described results satisfy the chosen criteria, i.e., they are composed of at least two and no more than six communities, each with a cardinality of at least two nodes. We have executed several supervised classification algorithms to evaluate how well they will perform on new data. To do this we treated the output of the unsupervised methods as the ground truth label, and through a k-fold procedure, we trained all the classification models with the same sample of data inside each fold, ensuring that performance differences among the algorithms were not due to chance. The results for all the classification models applied to each partition are summarized in Figure 2.7, which displays the partitions in rows and the classification models in columns. The values in Figure 2.7c represent the Balanced Accuracy (BA). A higher BA, marked by dark green, indicates better performance in predicting the outcome of the test data. Because the procedure involves k-fold validation, the values represented are the means of the 30 folds applied. At the end of each training with the k-based training set, the procedure computes the BA. After the last fold, the average accuracy is computed. Generally, every method performs very well with any partition except for the Neural Network, which achieves a maximum average accuracy slightly higher than 50%. Among the others, SVM is the one that generally achieves the highest performance, often exceeding 95%. All the partitions, especially those found through Fuzzy K-means, obtain interesting results, with BA higher than 90% in almost all supervised classifications (excluding NN). Figure 2.7a and Figure 2.7b represent the Macro F1 and Weighted F1 scores, respectively. The missing clustering methods and those with null values in these two figures have null recall and precision, i.e., the methods fail to assign the units of at least one cluster correctly. As for the accuracy, the support vector machine has the highest value of correct classification for all the partitions found with the clustering methods.

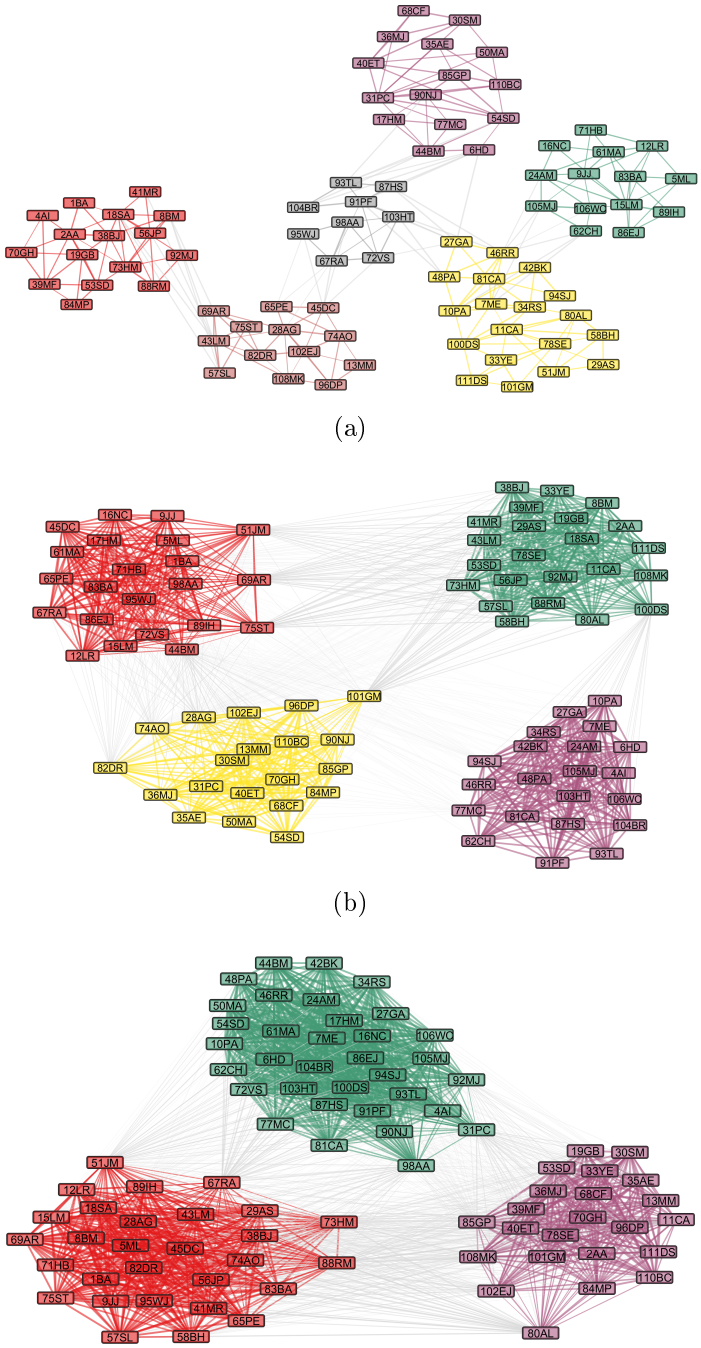


Figure 2.6: Networks and communities detected through (a) Edge Betweenness in a Planar Maximally Filtered Graph configuration, (b) Louvain and (c) Leading Eigen methods using a Leave-One-Out robustness metric based on positive correlation matrix.

Partition Comparison All validated partitions underwent pairwise analysis to examine whether some LAAs were clustered together in both partitions.

This comparative assessment aimed to discern similarities between different partitioning methodologies and parameter settings. As illustrated in Figure 2.8, five measures were employed to this scope.

Precision (Figure 2.8a) and Recall (Figure 2.8b) were utilised to compute both the F1 score (Figure 2.8c) and the Jaccard Index (Figure 2.8d). These metrics were calculated following the constructing of a co-membership confusion matrix for each pair of partitions. The F1 score reflects the effectiveness of the clustering methodologies in capturing shared LAAs. Indeed, it approaches values near 1, particularly for partitions derived from different Fuzzy K-means techniques and fuzzy parameter values, reflects the effectiveness of the clustering methodologies in capturing shared LAAs. Notably, F1= 1 was observed on the pairs composed of K-means Silhouette and various Fuzzy K-means specifications. Additionally, values exceeding 0.7 were obtained for comparison of partitions derived from Gaussian Mixture Models (GMM), albeit with variations based on assumptions regarding the covariance matrix. The Jaccard index results mirror those obtained through the F1 score.

Furthermore, Mutual Information (Figure 2.8e), Normalised Mutual Information (Figure 2.8f), and Adjusted Mutual Information (Figure 2.8g) were computed through entropy estimation. In general, all three measures exhibited lower values compared to the F1 score and the Jaccard index, reflecting the nuanced differences in partitioning outcomes and the level of information overlap between them.

It is noteworthy that both the K-means algorithm utilizing the Silhouette criterion and the Fuzzy K-means algorithm employing the Partition Entropy criterion with $m = 2.0$ yield identical partition. This alignment is considered a plausible optimal partition of the sample data into two distinct groups. Subsequently, we present the feature values of the LAA as shown in Figure 2.9. Group A comprises LAAs that are generally smaller than those in Group B. For example, the mean volume of appendages in Group A (5 cm^3) significantly differs ($p < 0.001$) from that of Group B (8.5 cm^3). Similarly, ostium area and perimeter, and the main path length support this conclusion, with average values of 6.5 cm, 2.94 cm^2, and 4.7 cm respectively for Group A, significantly differing ($p < 0.001$) from those of Group B (7.8 cm, 4.3 cm^2, and 6.0 cm). Tortuosity, curvature, and ostium-to-LAA area ratio are not significant. Eccentricity suggests that appendages in Group A possess a more rounded ostium than those in Group B ($p < 0.001$). Moreover, appendages in Group

B demonstrate a more intricate structure than those in Group A. Specifically, both the number of trabeculae ($p < 0.001$) and the complexity of the main path ($p < 0.001$) are higher in Group B, with an average of 12 trabeculae for appendages in Group B and 6 for those in Group A. Additionally, the relative complexity of the main path averages 0.33 in Group B and 0.25 in Group A. Table 2.5 presents the mean, median, and standard deviation for the two groups, while the appendages clustered together are show in Appendix B.

2.4 Discussions

This chapter laid the groundwork for establishing a feature-based group identity among LAA morphologies. Recent works in the field, particularly highlighted by the works of Wu et al. [158], Yaghi et al. [160], have elucidated the deficiencies inherent in the subjective classification methods employed since the seminal work of Wang et al. [152]. These studies highlight the importance of considering specific morphological characteristics, particularly in assessing clot-formation risk.

We have evolved the LAA classification beyond simplistic subjective assessments to embrace a more nuanced and empirical approach. While the contemporary understanding of LAA morphological classification still hinges upon subjective assessments, we proposed a data-driven approach through rigorous analysis and empirical data suitable to differentiate LAA identity groups. The many morphological features identified in Chapter 1, some of them highlighted by some authors as influencing clot-formation risk, are the core of our methodologies. Hence, the proposed approach encompasses a broad spectrum of morphological characteristics, including but not limited to appendage and ostium size.

Additional features that may be utilised to improve the classification include the minimum and maximum diameter of the ostium, lacunarity, fractal dimension, and the number of lobes. Incorporating the minimum and maximum diameters of the ostium can provide valuable information regarding the size and variability of the entrance to the LAA. Lacunarity, a measure of the gaps within a structure, can offer insights into the complexity and heterogeneity of the LAA morphology concerning trabeculations. The fractal dimension quantifies the intricate, self-similar patterns within the LAA, which can enhance understanding of its structural intricacies. Lastly, considering the number of

lobes can further refine the classification by detailing the LAA branching patterns and segmentation.

The seminal work of Wang et al. [152] catalysed LAA classification. Di Biase et al. [41], Kimura et al. [78], Li et al. [88], Słodowska et al. [138] reevaluated the classification by introducing refinement and differentiating the classes. However, such classification still relied heavily on subjective impressions, leading to inconsistencies and inaccuracies in LAA categorisation, as pointed out by Wu et al. [158], where the consensus among three clinicians was just 28% out of more than 2.000 patients, thus shedding light on the limitations of such subjective approaches. By contrast, a feature-based classification system offers a more objective and scientifically robust framework for delineating LAA classes.

We avoid the prevailing classification paradigm to redefine LAA morphology groups, opting for an elaborate exploration using diverse clustering methodologies. Acknowledging the limitations of the current classification system as the definitive label, we set out a systematic approach to uncover new groupings that better reflect the intricacies of LAA characteristics.

Central to our approach is the acknowledgement of the inherent ambiguity surrounding the number and composition of these groups. Factors such as cardinality add complexity to the clustering process, prompting us to employ various clustering methods, each harnessing multiple techniques. For instance, community detection in a Minimum Spanning Tree yields many clusters, while others, like Fuzzy K-means, tend to produce more compact groupings. Our criteria for selection prioritise clusters that strike a delicate balance between differentiation and simplicity, resulting in the retention of those with at least two but no more than six groups, each with a cardinality of at least two LAAs.

With the diverse outcomes generated by the clustering methods, identifying the most suitable partitions becomes paramount. To navigate this complexity, we adopt a systematic approach, treating each partition as a potential ground truth and subjecting them to a battery of supervised classifiers. We approve the ones with a maximum average accuracy higher than 70% through meticulous evaluation. It is worth noting that the Support Vector Machine emerges as the frontrunner in performance, displaying superior performance compared to alternatives such as Neural Network, which, in the adopted architecture, represents the worst classifier.

Having identified candidate partitions with commendable accuracy levels,

we examine their similarities using robust metrics like mutual information and the Jaccard index. These analyses offer valuable insights into the degree of overlap and distinction between various partitions. Notably, the partition identified through K-means and Fuzzy K-means result into the same patition, which is composed of two distinct groups, one with smaller appendages, few trabeculae, and round ostium, and vice versa.

In addition to delineating distinct LAA groups, our analysis unveils interesting patterns of overlap, with certain LAA pairs consistently appearing across multiple partitions. This observation suggests the interconnectedness of LAA characteristics.

Finally, the groups identified can not be easily comparable to the classifications proposed in the literature. Here, we compare our results with the methodologies that align most closely with our research objectives, specifically focusing on non-invasive procedures, such as coronary computed tomography angiography (CCTA) of MDCT, and employing precise measurements. Consequently, we have chosen not to compare our results to studies involving cadavers, those that consider the number of lobes, or those that focus on device positioning, as these aspects are outside the scope of our analysis.
For instance, works such as those by Su et al. [137], López-Mínguez et al. [93], Kaminski et al. [74], Wang et al. [150], Słodowska et al. [138], which primarily involve cadaver studies or focus on the number of lobes, have been excluded from our consideration. Studies like those by López-Mínguez et al. [93], Cressa et al. [34], which deal with LAA occlusion device positioning, have also been excluded. Furthermore, we also exclude a direct comparison with the method of Lacomis et al. [84], Koplay et al. [82]. Although they utilised CCTA and MDCT, respectively, their classification is based on the orientation and location of the LAA tip, avoiding a direct comparison with our method. These studies do not align with the imaging techniques or measurement parameters central to our classification framework.
Instead, we position our classification within studies employing non-invasive imaging techniques and emphasising quantitative measurements. Notable examples of the studies we consider relevant include Wang et al. [152], Li et al. [88], which used MDCT (the first two) and LAA angiography (the latter) for detailed measurements of the LAA. Although Wang et al. [152], Li et al. [88] have measured several features of the LAA, including ostium size, appendage curvature, and volume, their classification approach still necessitates human

intervention to assign an appendage to a specific group. Indeed, they found four and eleven LAA types, respectively. They assign the LAA to clusters by similarity shape, e.g., Chicken Wing, Windsock, Hook, Swan, Ring, ... This manual assignment introduces subjectivity and potential variability into the classification process as highlighted by Wu et al. [158], Yaghi et al. [160]. Instead, our approach categorised the LAAs based only on their quantitative morphological features, automatically computed without human intervention and without referring to similar shapes.

The classification proposed is morphologically based; hence, no patient or clinical data were considered. However, if clinical outcome data were available, such as the formation of blood clots, it would be possible to enhance the evaluation through clinical outcome metrics. By collecting longitudinal data on patient outcomes, short-term improvements could involve including patients with atrial fibrillation to identify characteristics associated with blood clot formation within the longitudinal study. In the long term, metrics such as Quality-Adjusted Life Years (QALY) and survival rates could be incorporated to validate the theoretical framework.

As we reflect on the implications of our findings, it becomes evident that the quality and robustness of the results could be further enhanced by increasing the sample size. We endeavour to contribute to a more deep understanding of LAA groupings through continued exploration and refinement.

2.5 Conclusions

In conclusion, this chapter laid the foundation for a novel approach to understanding LAA morphology through a feature-based grouping methodology. We embraced a data-driven approach, informed by recent works in the field that have highlighted the limitations of traditional subjective assessments.

By integrating various clustering methodologies and rigorous analysis of empirical data, we have proposed a more objective and scientifically robust framework for delineating LAA identity groups. This approach considers a broad spectrum of morphological characteristics which are hard to appreciate visually.

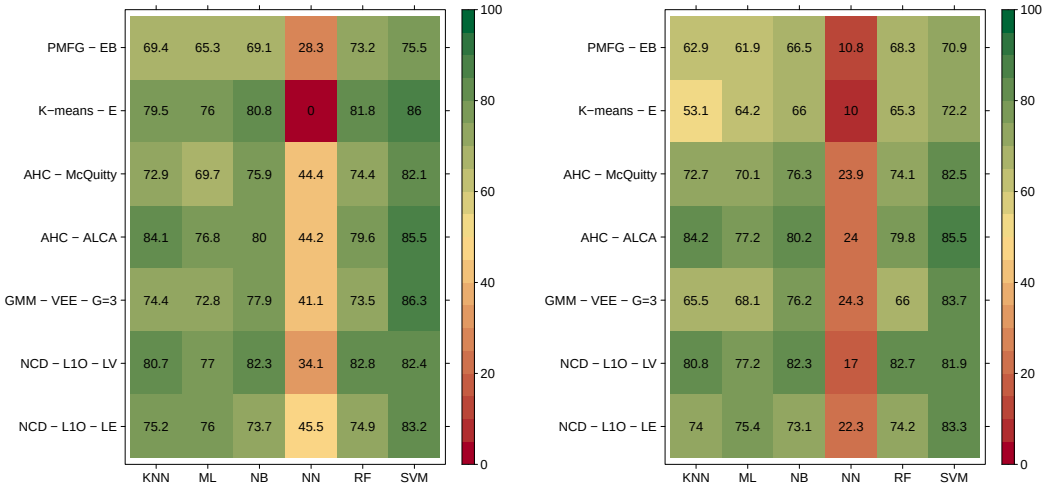
Although some clustering techniques, such as K-means and Fuzzy K-means, have revealed the presence of two distinct LAA groups, others have divided the sample into a more granular partition. This result suggests that the ap-

pendages can be easily categorised into two groups by significant morphological differences. Specifically, the LAAs can be categorised based on size and internal complexity variations, indicating a clear bifurcation in structural attributes. This differentiation is critical as it provides insights into the anatomical and possibly functional heterogeneity of LAAs, which could have implications for clinical assessments and interventions.
However, it is important to acknowledge that the limited sample size in the study has significantly influenced the determination of the optimal partitioning of these clusters. The small number of samples constrains the robustness of the clustering results, potentially leading to an incomplete understanding of the underlying morphological diversity. This limitation suggests that the current clusters may actually be composed of multiple subclusters that are not distinctly identified or partially identified by some methods due to insufficient data points.

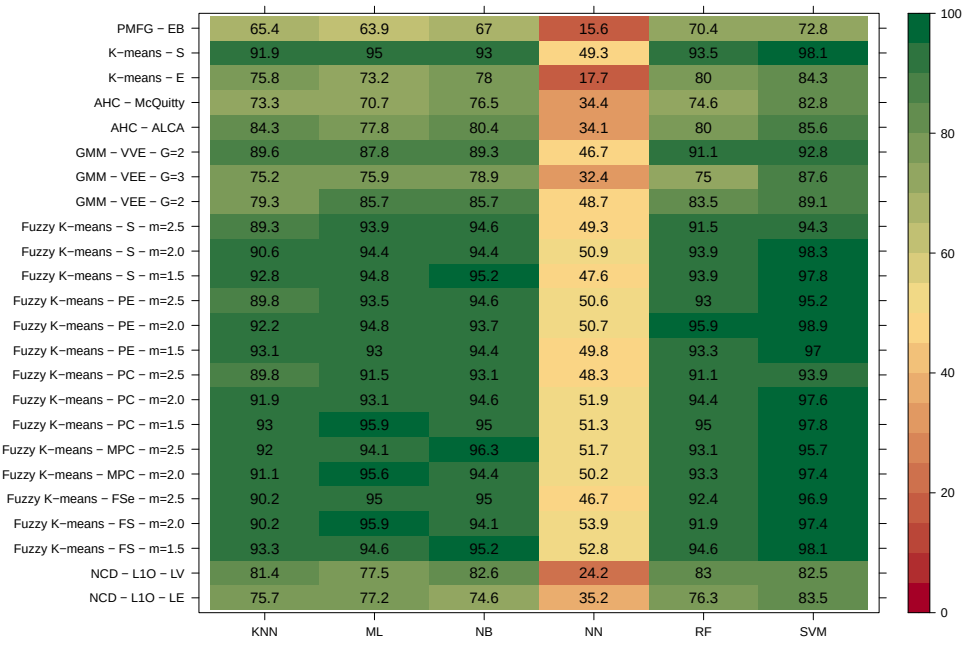
While the groups identified may not directly align with existing classifications in the literature, our approach offers a more comprehensive characterization based on multiple extracted LAA features. Future research endeavours should aim to expand the sample size further to enhance the quality and robustness of our results, contributing to a more profound understanding of LAA groupings. Moreover, exploring the clinical implications of these identified LAA groups would be beneficial, particularly concerning clot-formation risk and treatment strategies. This study represents a significant advancement in LAA morphology classification, offering a promising avenue for future research and clinical applications.

Table 2.4: Number of clusters for each partition found. Cluster cardinalities are in descending order, and no correspondences belong between different methodologies. The cluster cardinality column shows at most six clusters, if present. (n) indicates the presence of n groups of cardinality equal to 1. (*) indicates the presence of just one cluster with the entire data sample belonging to it.

Method	Technique (Parameters)	Number of Clusters	Cluster Cardinality
K-means	Elbow	6	$\{C_1 = 21, C_2 = 20, C_3 = 17, C_4 = 16, C_5 = 9, C_6 = 3\}$
	Gap Statistic	1 (*)	$\{C_1 = 86\}$
	Silhouette	2	$\{C_1 = 46, C_2 = 40\}$
Fuzzy K-means	Partition Coefficient ($m = 1.5$, $m = 2.0$)		
	Partition Entropy ($m = 1.5$, $m = 2.0$)		
	Modified Partition Coefficient ($m = 2.0$)		
	Silhouette ($m = 1.5$, $m = 2.0$)		
	Fuzzy Silhouette ($m = 1.5$, $m = 2.0$)		
	Partition Coefficient ($m = 2.5$)		$\{C_1 = 45, C_2 = 41\}$
	Partition Entropy ($m = 2.5$)		
	Silhouette ($m = 2.5$)		
	Fuzzy Silhouette ($m = 2.5$)		
	Modified Partition Coefficient ($m = 2.5$)		
	Modified Partition Coefficient ($m = 1.5$)	20 (1)	$\{C_1 = 10, C_2 = 7, C_{3,4,5,6} = 6, \dots\}$
Hierarchical Clustering	Agglomerative (Average, McQuitty)	3	$\{C_1 = 35, C_2 = 26, C_3 = 25\}$
	Agglomerative (Complete)	13	$\{C_1 = 11, C_{2,3,4} = 8, C_{5,6} = 7, \dots\}$
	Divisive (Diana)	14	$\{C_{1,2} = 11, C_3 = 10, C_4 = 7, C_{5,6} = 6, \dots\}$
Gaussian Mixture Model	VVE	2	$\{C_1 = 47, C_2 = 39\}$
	VEE		
		3	$\{C_1 = 39, C_2 = 38, C_3 = 9\}$
Complete Network	Edge Betweenness	1	$\{C_1 = 86\}$
	InfoMap		
	Leiden	3	$\{C_1 = 33, C_2 = 30, C_3 = 23\}$
	Louvain		$\{C_1 = 30, C_2 = 29, C_3 = 27\}$
	Leading Eigen		$\{C_1 = 43, C_2 = 26, C_3 = 17\}$
Minimum Spanning Tree	CLCA - InfoMap	19	$\{C_{1,2,3,4} = 6, C_{5,6} = 5, \dots\}$
	ALCA - InfoMap	17	$\{C_1 = 8, C_3 = 7, C_{3,4,5,6} = 6, \dots\}$
	SLCA - InfoMap	16	$\{C_1 = 7, C_{2,3,4,5,6} = 6, \dots\}$
	CLCA - Leading Eigen	14	$\{C_{1,2} = 11, C_3 = 10, C_{4,5} = 9, C_6 = 8, \dots\}$
	CLCA - Louvain	12	$\{C_{1,2} = 9, C_{3,4,5,6} = 8, \dots\}$
	ALCA - Leading Eigen		$\{C_1 = 12, C_{2,3} = 10, C_{4,5} = 9, C_6 = 8, \dots\}$
	SLCA - Leading Eigen	12 (1)	$\{C_1 = 12, C_2 = 11, C_3 = 10, C_4 = 9, C_{5,6} = 8, \dots\}$
	ALCA - Leiden	11	$\{C_{1,2,3,4} = 10, C_5 = 9, C_6 = 8, \dots\}$
	CLCA - Leiden		$\{C_1 = 11, C_2 = 9, C_{3,4,5,6} = 8, \dots\}$
	SLCA - Edge Betweenness	9	$\{C_1 = 13, C_{2,3} = 12, C_{4,5,6} = 9, \dots\}$
	SLCA - Leiden		$\{C_1 = 12, C_{2,3} = 11, C_4 = 10, C_{5,6} = 9, \dots\}$
	SLCA - Louvain		$\{C_{1,2} = 12, C_{3,4} = 11, C_5 = 10, C_6 = 9, \dots\}$
	ALCA - Edge Betweenness		$\{C_{1,2} = 12, C_{3,4} = 11, C_{5,6} = 9, \dots\}$
	ALCA - Louvain		$\{C_{1,2} = 12, C_{3,4,5} = 10, C_6 = 9, \dots\}$
	CLCA - Edge Betweenness		$\{C_1 = 14, C_2 = 12, C_{3,4} = 10, C_5 = 9, C_6 = 8, \dots\}$
Planar Maximally Filtered Graph	Edge Betweenness	6	$\{C_1 = 19, C_2 = 16, C_3 = 15, C_4 = 14, C_5 = 13, C_6 = 9\}$
	InfoMap	8	$\{C_1 = 14, C_{2,3,4} = 12, C_5 = 10, C_6 = 9, \dots\}$
	Leiden	7	$\{C_1 = 16, C_2 = 14, C_3 = 13, C_{4,5} = 12, C_6 = 10, \dots\}$
	Louvain		$\{C_1 = 21, C_2 = 13, C_{3,4} = 12, C_5 = 10, C_6 = 9, \dots\}$
	Leading Eigen		$\{C_1 = 17, C_2 = 16, C_3 = 14, C_4 = 13, C_5 = 12, C_6 = 9, \dots\}$
Positive Correlated Network	Edge Betweenness	1 (*)	$\{C_1 = 86\}$
	InfoMap		
	Leiden	5	$\{C_1 = 23, C_2 = 22, C_3 = 20, C_4 = 13, C_5 = 8\}$
	Louvain	4	$\{C_1 = 24, C_2 = 22, C_3 = 21, C_4 = 19\}$
	Leading Eigen	3	$\{C_1 = 34, C_2 = 29, C_3 = 23\}$
Leave-one-Out Robustness Network	Edge Betweenness	1 (*)	$\{C_1 = 86\}$
	InfoMap		
	Leiden	4	$\{C_1 = 24, C_2 = 24, C_3 = 21, C_4 = 17\}$
	Louvain		$\{C_1 = 23, C_2 = 22, C_3 = 21, C_4 = 20\}$
	Leading Eigen	3	$\{C_1 = 34, C_2 = 29, C_3 = 23\}$
K-fold CV Robustness Network	Edge Betweenness	1 (*)	$\{C_1 = 86\}$
	InfoMap		
	Leiden	5 (1)	$\{C_1 = 30, C_2 = 21, C_3 = 17, C_4 = 17, C_5 = 1\}$
	Louvain	4	$\{C_1 = 25, C_2 = 22, C_3 = 20, C_4 = 19\}$
	Leading Eigen	2	$\{C_1 = 47, C_2 = 39\}$

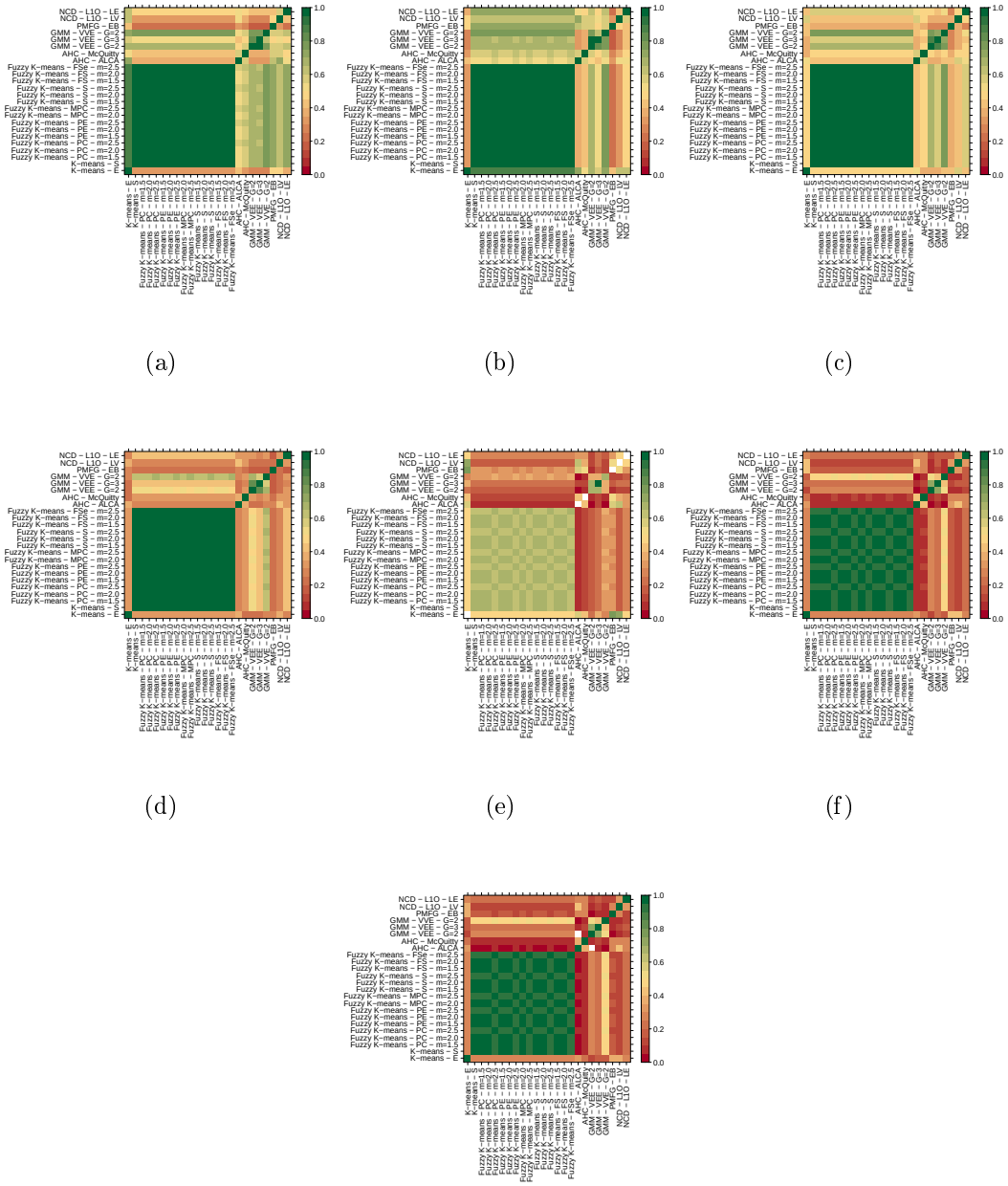


(a) (b)



(c)

Figure 2.7: Graphical representation of the Macro F1 score (see Figure 2.7a), Weighted F1 (see Figure 2.7b), and average Balanced Accuracy (see Figure 2.7c) obtained during the K-fold procedure to evaluate the reliability of partitions. The partitions are depicted on the y-axis, while the x-axis illustrates the supervised classification algorithm. A higher value of the indices indicates superior performance in classifying new units. Clustering techniques with a null value in either Macro F1 score or Weighted F1 were omitted to enhance graphical representability.



(a) (b) (c)

(d) (e) (f)

(g)

Figure 2.8: Similarity measures for partition comparison. (a) Precision, (b) Recall, (c) F1 score, (d) Jaccard index, (e) Mutual Information, (f) Normalised Mutual Information, (g) Adjusted Mutual Information.

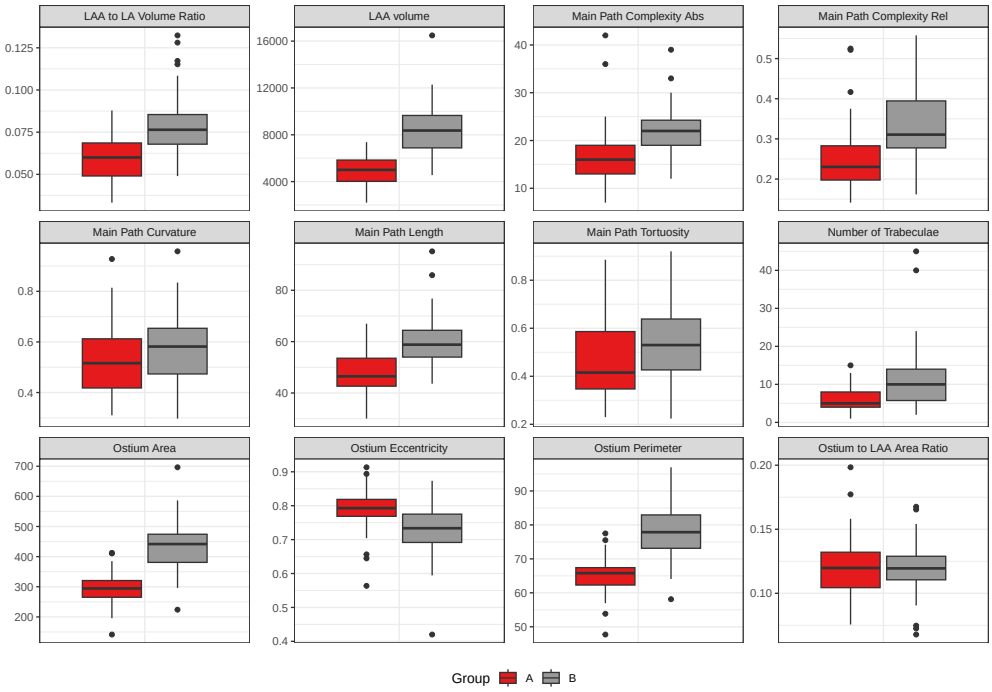


Figure 2.9: Boxplot by group of the features.

Table 2.5: Main statistics of the features by group.

Feature	Unit	Group	Mean	Median	St.Dev.
Ostium Perimeter	cm	A	6.52	6.58	0.56
		B	7.78	7.79	0.76
Ostium Area	cm^2	A	2.94	2.94	0.53
		B	4.31	4.42	0.84
Ostium Eccentricity		A	0.78	0.79	0.06
		B	0.73	0.73	0.08
Ostium Area to LAA Area Ratio		A	0.12	0.12	0.02
		B	0.12	0.12	0.02
LAA Volume	cm^3	A	4.96	5.00	1.21
		B	8.48	8.36	2.28
LAA to LA Volume Ratio		A	0.06	0.06	0.01
		B	0.08	0.08	0.02
Number of Trabeculae		A	5.98	5.00	3.34
		B	11.78	10.00	9.07
Length of the Main Path	cm	A	4.73	4.65	0.81
		B	6.01	5.88	1.06
Absolute Complexity of the Main Path		A	16.72	16.00	6.24
		B	22.35	22.00	5.10
Relative Complexity of the Main Path		A	0.25	0.23	0.08
		B	0.33	0.31	0.10
Tortuosity of the Main Path		A	0.46	0.42	0.18
		B	0.54	0.53	0.18
Curvature of the Main Path		A	0.52	0.52	0.15
		B	0.56	0.58	0.15

Chapter 3

MicroRNA and Messenger RNA Interaction Analysis

MicroRNAs (miRNAs) regulate gene expression by binding to target messenger RNA (mRNA) molecules. This regulation is fundamental to various cellular processes, including development, differentiation, and other essential cell functions. Since a single miRNA can target multiple mRNAs and vice versa, miRNA-mRNA interactions are complex and can be effectively represented as a bipartite network. This chapter delves into the intricate world of miRNA-mRNA interactions, highlighting the importance of network analysis and competitive modelling in understanding these complex relationships.

The first section of this chapter provides a biological introduction to miRNAs and mRNAs. It covers the molecular characteristics of these entities and describes the experimental techniques used to measure gene expression and validate miRNA interactions. Techniques such as quantitative polymerase chain reaction (PCR), microarrays, and next-generation sequencing, are commonly employed to quantify miRNA and mRNA levels.
The concept of bipartite networks is also introduced, providing a foundation for understanding the miRNA-mRNA interaction models. The section also outlines the strategy for developing a miRNA-mRNA competitive model, which involves identifying and analysing the competing interactions between miRNAs and their target mRNAs.

In the second section, we present a detailed miRNA-mRNA competing model based on bipartite network theory. This model includes a pre-procedure aiming at normalising the data followed by a two-phase extraction procedure, where the first one is based on miRNA and mRNA expression data and the

other on an interaction score. While the expression data are derived from in-vitro experiments, the interaction scores are from computational predictions and experimental validations. This dual-phase approach enables a robust identification of miRNA-mRNA interactions, accounting for expression dynamics and binding affinities. The bipartite network model captures the complexity of these interactions, allowing for the visualisation and analysis of competing miRNA-mRNA relationships. The validation process includes comparing model predictions with experimentally confirmed interactions to assess the model's performance.

In the third section we apply the proposed miRNA-mRNA competing model using data from *in vitro* experiments. These experiments typically involve cell cultures treated under various conditions to observe changes in miRNA and mRNA expression. By applying the model to *in vitro* data, we can test its predictive power and accuracy in identifying true miRNA-mRNA interactions.

The final section concludes by summarising the key findings and implications of the miRNA-mRNA competing model. It emphasises the importance of integrating network analysis and competitive modelling in miRNA research.

Abstract

This study explores miRNA-mRNA interactions, emphasizing the importance of network analysis and competitive modelling. Due to the high costs associated with RNA-induced silencing complex, microRNA target identification has increasingly adopted a computational approach. MicroRNAs are a particular type of regulatory RNA that plays in RNA interference, a cell process where small RNA molecules physically interact with mRNA molecules to suppress gene expression. Here, we present a competing model using bipartite network theory that involves data normalization and a two-phase extraction process based on expression data and interaction scores. This approach effectively identifies miRNA-mRNA interactions. The model application to *in vitro* data tests its predictive accuracy. The study concludes that integrating network analysis and competitive modelling is crucial in miRNA research. Among the evaluated models, the product sampling (PS) model with PITA scores performed best, highlighting the importance of combining binding scores with expression data.

3.1 Biological Foundations and Experimental Techniques

MicroRNAs were initially identified as a peculiar gene expression regulation mechanism specific to the nematode *Caenorhabditis elegans* [156, 130]. However, subsequent research has established miRNAs as critical regulators of gene expression in eukaryotes, involved in many cellular processes. This regulatory role is part of the broader RNA interference (RNAi) mechanism, encompassing various small RNA-mediated pathways that influence gene expression post-transcriptionally [47]. Dysregulation of RNAi, including miRNA activity, is associated with numerous human diseases, ranging from cancers to neurological disorders [46, 23].

A biological explanation of miRNA activity precedes the description of state-of-art methodologies, which partially account for the goal of this chapter. Hence, a brief introduction of basic genetic concepts is reported before going into detail about miRNA-mRNA literature. The description of the miRNA target prediction procedures follows, such as the miRNA-mRNA correlation network and competitive model.

3.1.1 Molecular Characterization of Key Biological Entities

Understanding the fundamental biological processes and molecules involved in gene regulation is essential for comprehending the complex interactions within cellular systems. The phases that lead the genetic information from DNA to protein synthesis can be summarised as follows:

1 **DNA replication**: It is a vital process during cell division in which DNA produces an exact copy of itself. Structurally, DNA is formed by a double helix of two complementary strands. These strands are made up of nucleotides, each characterised by a nitrogenous base - cytosine (C), guanine (G), adenine (A), or thymine (T). In canonical Watson-Crick base pairing, A forms a base pair with T, and G forms a base pair with C, thus dictating the complementary nature of the DNA strands. During replication, the two strands of the DNA helix separate, each serving as a template for synthesising a new complementary strand. This synthesis is facilitated by a molecular complex containing the enzyme

DNA polymerase, which ensures the correct pairing and binding of the complementary nucleotides to form the new DNA strands.

2 **Transcription**: it is the biological process by which genetic information encoded in DNA is transcribed into mRNA. This process is facilitated by a molecular complex containing the enzyme RNA polymerase, which synthesises mRNA by using one of the DNA strands as a template. The resulting RNA strand is complementary to the DNA template strand, with one key difference: instead of thymine (T), adenine (A) pairs with uracil (U) in the RNA molecule.

3 **Translation**: mature mRNA molecules exit the nucleus after the transcription process and enter the cytoplasm, where they encounter ribosomes. The ribosomes read the sequence of codons in the mRNA and use this sequence as a template to assemble a chain of amino acids, ultimately forming a protein.

A triplet of close nucleotides in an mRNA forms a codon. Among the 64 possible codons, 61 codify an amino acid, while three serve as stop signals interrupting the polypeptidic chain assembly. While each of the 61 codons corresponds to a specific amino acid, an amino acid can correspond to several codons. For instance, the nucleotide UCU corresponds to the Serine (Ser/S) amino acid, but Ser/S correspond to UCU, UCC, UCA, and UCG.

There are also DNA strands that do not have any coding role. Although they were considered as *junk DNA* in the past, many recent studies emphasised the importance of non-coding DNA for cell life and evolution [114]. As (coding) DNA is transcribed into (coding) RNA, non-coding DNA is transcribed into non-coding RNAs. Different types of non-coding RNAs exist, such as transfer RNA, ribosomal RNA, and regulatory RNA, each with its function.
MicroRNAs are a particular type of regulatory RNA that plays a role in RNA interference [60], which is a cell process where small RNA molecules physically interact with mRNA molecules to suppress gene expression. The activity of RNAi molecules relies on forming a protein complex known as the RNA-induced silencing complex (RISC). The RISC integrates one strand of the RNAi molecules, which serves as a guide to identify and bind to complementary mRNA transcripts. Although the composition of the RISC complex varies among different organisms, it consistently includes a protein family called Argonaute. RNAi is utilised in numerous experiments by introducing synthetic

double-stranded RNAs into cells to suppress specific genes selectively. This technique aids in identifying the components essential for cellular processes.

3.1.2 Computational Approaches to miRNA Target Prediction

One of the most prevalent experimental strategies for investigating miRNA targets involves the immunoprecipitation of RISC proteins [30, 67]. Due to the high costs associated with such experiments, miRNA target identification has increasingly adopted a computational approach, replicating biological experiments through simulations. In *in silico* predictions of molecular binding interactions typically guide the selection of significant *in vitro* experiments.
MicroRNA binding prediction algorithms rely on Watson-Crick (WC) base-pair matching [151, 162, 52]. Although perfect complementarity between miRNA and mRNA pairs is rare, even a six-base-pair match can be sufficient to suppress gene expression. Some methods also incorporate miRNA expression profiles to predict miRNA targets [104, 33, 32]. Some tools identify nucleotide complementarity by analysing RNA sequences and assigning a binding score to the miRNA-mRNA pair. These tools are:

- **TargetScan** [101]: it detects miRNA targets through base-pairing interactions, focusing on nucleotides 2-8 of the miRNA (the seed region) and basing its target prediction on perfect WC seed complementarity, referred to as canonical matches. TargetScan distinguishes four types of canonical binding sites, ranging from 6mer (weakest) to 8mer (strongest) matches as the number of matching nucleotides. The presence of the A base at the first position (the A anchor) increases the specificity of target prediction; hence, the A conservation characterises stronger matching. TargetScan excludes non-canonical matches as they are the minority.

- **miRanda** [44]: it identifies potential miRNA binding sites in genomic sequences, including non-canonical matches. Target search employs a complementary assessment to evaluate if two sequences are complementary, computes a matching score, and associates significant matches with the energy of physical interaction.

- **PITA** [77]: it accounts for the energy cost of making the target accessible for miRNA binding. Scores are computed as the difference between

the free energy gained from miRNA-target duplex formation and the energetic cost of unpairing the target. Lower PITA scores indicate higher probabilities of interaction.

3.1.3 Gene Expression Analysis

Modern genetic research uses gene expression analysis to investigate cellular activity and functions. The expression level of a gene corresponds to the amount of RNA transcript of that gene. The primary techniques aimed at quantifying gene expression are:

- **DNA microarray** [90] consists of microscopic DNA spots attached to a solid surface, each containing probes for specific DNA sequences. These probes bind to complementary DNA (cDNA) synthesised from RNA samples. The mRNA is collected from healthy and diseased individuals and converted into cDNA using reverse transcriptase, with each sample marked by distinct fluorescent dyes. The two samples are mixed and allowed to bind the microarray slide, a hybridisation process. Finally, microarray data are obtained by quantifying fluorescent pixel intensities. The described approach is named glass slide microarray, one of the available microarray approaches.

- **Next-Generation Sequencing**, NGS [61], is a set of techniques for massive genome sequencing and expression quantification, also used for genome comparisons, identifying gene variants, and discovering new miRNAs. NGS is more flexible than hybridisation-based methods because it allows to discover novel RNAs. NGS expression levels correlate highly with real-time PCR experiments.

- **Real-Time Quantitative PCR**, RT-PCR [28], is the standard technique for measuring gene expression, validating other profiling approaches, and diagnosing disease. The first step of the procedure consists of using reverse transcriptase to synthesise RNA to cDNA. Finally, cDNA is amplified via PCR, with quantification done in real-time using fluorescent dyes, where the fluorescence signal increases with the amount of replicated DNA.

Another powerful molecular biology technique that can be used to study gene expression is RNA Immunoprecipitation combined with microarray anal-

ysis (RIP-Chip) [54]. This approach isolates specific protein-RNA complexes using antibodies that recognize and bind to RNA-binding proteins. During the RIP-Chip process, cells are first lysed to release their contents, and an antibody specific to the RNA-binding protein of interest is used to precipitate the protein along with its bound RNA molecules. The immunoprecipitated RNAs are then purified and identified using microarray analysis, which involves hybridizing the RNAs to a microarray chip containing probes for thousands of genes. This provides a precise snapshot of the RNA molecules interacting with the protein of interest at a given time. Three quantities can be defined with such techniques:

- **Input sample**, IN: it is the RNA extracted from whole cells lysate.
- **Immunoprecipitated**, IP: the RNA molecule bound by a particular protein during the experiment.
- **Flow-through**, FT: the RNA extracted from the residual fraction of cells lysate after the immunoprecipitation experiment.

3.1.4 Computational Methodologies in Gene Expression Regulation

Regulatory network reconstruction has traditionally focused on transcriptional regulation, but integrating miRNAs into these networks is becoming increasingly important due to their role in post-transcriptional gene regulation. Despite their potential, few tools have effectively combined mRNA and miRNA expression data to infer regulatory interactions.

Several existing tools attempt to bridge this gap, albeit with limitations. For instance, MMIA [107] and MAGIA [126] utilise association metrics such as correlation and mutual information to infer regulatory interactions. However, MMIA only considers miRNAs that are significantly up or downregulated, excluding those that might be significantly correlated with their targets but do not meet specific differential expression thresholds. This approach limits the analysis to control/disease contrasts and excludes time-series data. GenMir++ [70] is another advanced algorithm based on a Bayesian model to identify miRNA targets from sequence features and expression data. However, GenMir++ suffers from computational inefficiency with large datasets and does not consider supporting information such as transcriptional regulation.

A common drawback of these tools is their failure to integrate the entire set of transcription factors in the network construction. Moreover, they often overlook network motifs, such as feedback and feed-forward loops, which are known to play significant roles in cancer development and other diseases.

MirConnX [69, 149] is a tool designed to address these needs. Unlike existing tools, mirConnX maximally exploits information from mRNA and miRNA expression data and integrates it with pre-compiled data on miRNA targeting and TF binding. This integration allows for the generation of condition-specific, genome-wide regulatory networks that reflect the unique characteristics of the user-supplied data, guided by prior knowledge.

3.1.5 Regulatory Networks of RNA

Competing endogenous RNAs (ceRNAs) play a crucial role in regulating RNA transcripts by competing for shared microRNAs [3]. Various models illustrate ceRNA regulation, showing how fluctuations in the levels of one or more miRNA targets impact the availability of miRNAs, thereby influencing the expression of other miRNA-regulated genes. These models range from describing direct interactions between targets and miRNAs to more complex scenarios involving multiple RNA targets and miRNAs. Some models emphasize the interaction between mRNA 3' UTRs and miRNAs, where the 3' UTRs are crucial in regulating gene expression. They often contain complementary sequences to specific miRNAs, allowing them to bind precisely to their target mRNAs, thus usually resulting in gene silencing. Finally, other models consider the roles of long non-coding RNAs. Despite substantial research, the detailed biochemical mechanisms of ceRNA regulation remain an active area of scientific inquiry.

The influence of ceRNA regulation has been documented in numerous studies, particularly concerning its roles in both normal cellular function and various diseases. However, the scientific community continues to explore and debate the full extent of ceRNA interactions and their implications.

The concept of miRNA "sponge" RNAs emerged as a strategy to induce sustained miRNA loss of function in cells [39]. Sponge RNAs are designed with binding sites complementary to specific miRNAs and expressed from transgenes. These binding sites typically target the miRNA seed region, enabling the sponge RNAs to sequester and inhibit a family of related miRNAs. This method was inspired by natural processes that inhibit miRNA activity within cells.

Artificial sponge RNAs contain multiple miRNA binding sites complementary to mature miRNA seed sequences and are usually driven by strong viral promoters, such as those from Simian Virus 40 or Cytomegalovirus. These sponge RNAs can be introduced into cells through transient plasmid transfection or stable viral transduction, and they have been widely utilized *in vitro* and preclinical research settings.

Over the past decade, methodologies in molecular biology have been profoundly influenced by network theory [157]. From a systems biology perspective, the focus is on the physical interactions between biological elements, collectively called the Interactome [35]. The Interactome encompasses the complete interactions between cellular components, representing a significant and ambitious challenge in molecular biology. Network models are ideally suited for representing and analyzing these molecular interactions, delineating the interactions and pathways that define various cellular processes.

Many molecular networks have been defined to capture different types of biological interactions. For instance, using bipartite representation, metabolic networks [71] represent the relationships between chemical reactions and their substrates. Protein-protein interaction networks [128] represent protein binding interactions. Gene regulatory networks [75] describe the mechanisms governing gene expression regulation through graphical models. These networks share several properties with macro-scale systems, such as scale-free degree distribution and preferential attachment behaviour, indicative of their underlying complexity and robustness.

MicroRNAs (miRNAs) and their targets form a bipartite network. However, the full extent of miRNA connectivity remains largely unknown due to the high costs associated with experimental procedures that limit the knowledge of specific binding interactions. In *in silico* predictions of miRNA, interactions offer a valuable approach to uncover molecular interactions that are not yet known.

3.2 Methods

In this section, we present a novel algorithm to evaluate miRNA interaction. The algorithm is composed of three phases: the pre-processing of data, the competitive interaction model, and the evaluation procedure.

3.2.1 Data Preparation

A pre-processing step is necessary to ensure data is suitable for analysis and model building. Converting microarray data into *in silico* data may be challenging, especially when dealing with the expression of molecules. Let E represent the experiment under investigation. Data derived from microarrays are quantified as pixel intensities of fluorescent dyes. The composition of the starting solution (or input), IN, the imhuno precipitated, IP, and flow-through solution, FT, are computed accordingly. The given data encompass the intensities of gene and miRNA expressions across the three phases of the experiment, which are represented by floor numbers. In *in silico* analysis, expression levels must be discrete, i.e., expressed by natural numbers, since the number of molecules cannot be fractional. Here, we present a technique used to convert fluorescent intensities from microarray output data into model input data. Our method focuses on three key points: correction for an error term, insurance of the additive property, and float-to-integer conversion.

Error Term Correction Microarray data are associated with an error term, ER. It is appropriate to correct data by its error to ensure model accuracy. We use the error term as a control measure to check wheter the expression is valid. In particular, if the expression of a molecule is smaller than its error term, than the molecule is unvalidated by set its expression to zero. Hence,

$$
\begin{aligned}
E_{IN} &= \begin{cases} E_{IN} & \text{if } E_{IN} > E_{IN}^{ER} \\ 0 & \text{otherwise} \end{cases} \\
E_{IP} &= \begin{cases} E_{IP} & \text{if } E_{IP} > E_{IP}^{ER} \\ 0 & \text{otherwise} \end{cases} \\
E_{FT} &= \begin{cases} E_{FT} & \text{if } E_{FT} > E_{FT}^{ER} \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
\tag{3.1}
$$

Additive Property The method described in the following section requires that the whole amount of molecules the system is composed of at the experiment beginning must equal the amount of molecules in the outputs. Particularly, as the experiment presents E_{IN} molecules at the beginning, the sum between the two parts that compose the output should ensure the same amount, i.e. E_{IP} and E_{FT}. We keep fixed the expression of the input E_{IN} and adjust

the expression of the outputs by assuming the input as a linear combination of them $E_{IN} = \alpha E_{IP} + \beta E_{FT}$. In particular, α and β are estimated through linear regression without constant term. The first term of the right-hand side of the equation, αE_{IP}, will be used to correct the expression of IP. In contrast, the left-hand side, βE_{FT}, is needed to estimate the parameters α. The expression of FT will be set as the difference between the ones of IN and IP to ensure the additive property. The expressions are adjusted as

$$\begin{aligned} E_{IN} &= E_{IN} \\ E_{IP} &= \min\{E_{IN}, \alpha E_{IP}\} \\ E_{FT} &= E_{IN} - E_{IP} = E_{IN} - \min\{E_{IN}, \alpha E_{IP}\} \end{aligned} \tag{3.2}$$

Expressions are defined in the domain of positive real numbers. Since our method interprets expressions as the number of molecules in the system, it is necessary to represent them using natural numbers. Therefore, converting floating-point numbers to integers is essential. We round the floating-point number to the nearest integer, providing a balance between overestimation and underestimation. Hence

$$\begin{aligned} E_{IN} &= [E_{IN}] \\ E_{IP} &= [E_{IP}] \\ E_{FT} &= [E_{FT}] \end{aligned} \tag{3.3}$$

3.2.2 Scenario Investigation

The previously described correction procedures facilitate the representation of gene and miRNA expressions within the positive discrete domain, denoted as $\mathbb{N}_0^+$. These procedures ensure consistency in data across the three experimental phases, maintaining both the expressions' additive property and non-negative expressivity. Despite these corrections, microarray analyses are conducted separately for genes and miRNAs, rendering the two sample sets independent. Consequently, while we can ascertain the composition of miRNA or gene samples individually, we lack information on the combined composition of these molecules within the same sample. This limitation prevents us from knowing the total quantity of each molecule in the system.

To address this issue, we explore a wide range of scenarios. Specifically, we recognise that the optimal composition of the system resides within the space $\mathbb{N}_0^{+2}$, and we investigate scenarios where the cardinality of molecules in

the system has a power of 10 to explore various order of magnitude. Although this set of scenarios is discrete and limited, it represents a concise way for exploring various system compositions, enabling us to identify the one that most closely approximates the real system. The scenario space, shown in Figure 3.1, illustrates the analysed scenarios.

For clarity, we designate each scenario in the space as $\mathcal{S}_G^M$, where G (M) represents the total number of genes (miRNAs) in the system. For example, $\mathcal{S}_5^7$ refers to the scenario in which the system comprises $100,000$ genes and $10,000,000$ miRNAs. Scenarios where $G = M$, represented along the main diagonal of the scenario matrix, are considered balanced. Conversely, scenarios in the upper triangular matrix contain a higher expression of miRNAs than genes. In comparison, those in the lower triangular matrix reflect a greater expression of genes relative to miRNAs.

3.2.3 MicroRNA-mRNA bipartite network

MicroRNA binding depends not only on the accessibility of mRNA binding sites but also on the total amount of the molecules in the cell. The method presented here considers the binding scores and how molecules influence each other. Indeed, the amount of a single miRNA can influence the binding of the other miRNAs by acting as a competitor or as a heckler. In our model, molecule amounts represent the initial system condition. In particular, we assume that genes and miRNAs come from two separate chambers, which, from now on, are called urns. When a molecule is selected from its urn, it may search for a compatible counterpart and try to bind. Of course, the higher the amount of a given molecule in its urn, the higher the probability of being selected. The binding probability of selected molecules is based on the binding scores, i.e., on the molecules' compatibility. The higher the binding scores, the higher the probability the two molecules bind. Once the molecules in the system cannot bind anymore, the system is refilled with new molecules, which may or may not comprise the previously unbound ones. In other words, if some molecules do not bind in a given moment, they can be selected again to bind in successive iterations. The procedure ends when no remaining molecules can bind together or after a given stop criteria.
The algorithm is based on the following steps:

0. Object initialisation:

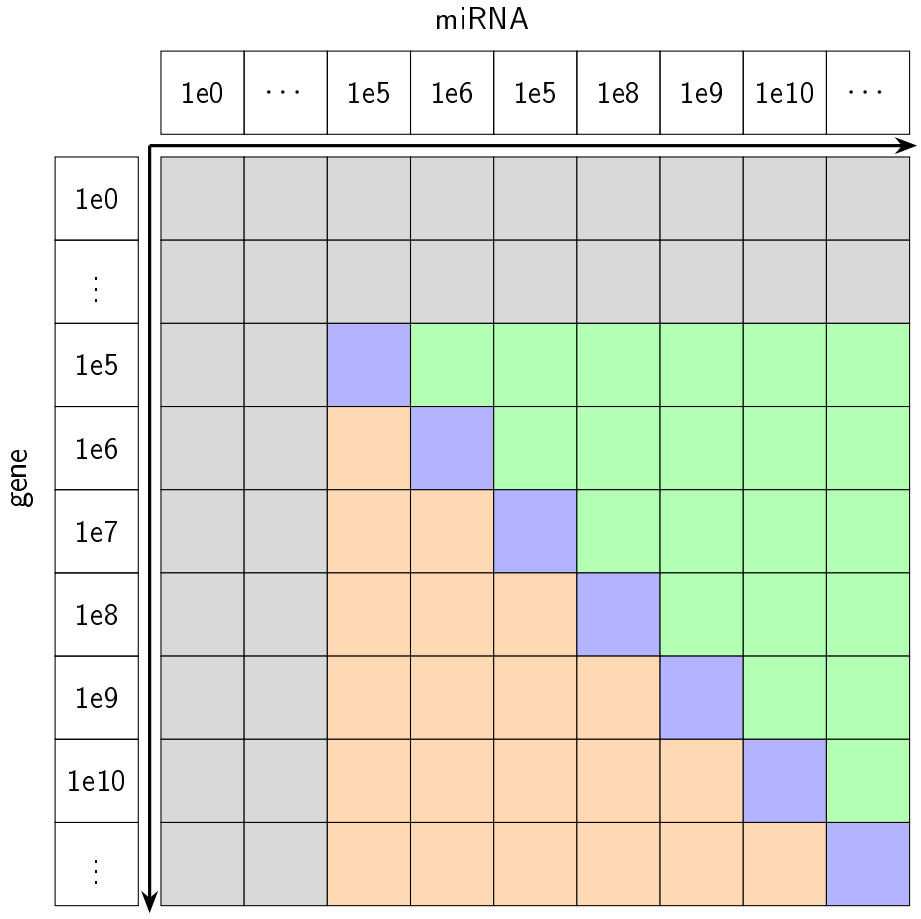


Figure 3.1: Scenario space. Gray cells represent scenarios that do not make sense to analyse because the total expression is too small. Blue cells refer to balanced scenarios, where the cardinality of gene and miRNA urns are equal. Green (orange) cells refer to unbalanced scenarios where the number of miRNA (gene) is greater than the gene (miRNA) one.

- An empty bipartite network, k, composed of no edges, $E(k)$, and no nodes, $N(k)$.
- miRNA urn U_M composed of the total amount of input expression E^M_{IN}.
- mRNA urn U_G composed of the total amount of input expression E^G_{IN}.

1. p miRNAs and q mRNAs are extracted from the two urns independently. The extraction probability is proportional to the expression of the miRNAs and mRNAs in the urns, i.e. $p(m \in M) = \frac{E^m_{IN}}{\sum_m E^m_{IN}}$ and $p(g \in G) = \frac{E^g_{IN}}{\sum_g E^g_{IN}}$.

2. Search for all the plausible links between the molecules in p and q, and associate the corresponding binding score, $L = \{(p-q) : s_{pq}\}$, chosen between the available tools (TargetScan, miRanda, PITA, ...).

3. The pairs are extracted using the binding scores as a probability until no more molecules in q and p can bind together, i.e., the pair extraction stops when the system is composed of incompatible molecules.

4. The pairs extracted in step 3., L^*, feed the bipartite network k such that $E(k)$ is augmented with the new pairs and $N(k)$ is fostered with the new pairs' molecules.

5. Check convergence. If one of the two urns is empty or a stop criteria is reached, STOP. Jump to step 1. otherwise.

3.2.4 Model evaluation

The bipartite network k that is the output of our model is the corresponding composition of the immune precipitated expression, E_{IP}, of the microarray, i.e. the bound molecules. A robust *in silico* model should reflect the microarray data. To compare the output of our model with the microarray data, we use several metrics to compute the model's overall goodness by evaluating the overlapping information. Let $P(g^{'})$ be the probability mass function of mRNAs in k and $Q(g)$ be the probability mass function of mRNAs in E_{IP}, then we compute:

- **Kullback-Leibler Divergence**, DKL: it measures the difference between two probability distributions. It is defined as

$$DKL(P||Q) = \sum_i P(i) log\left(\frac{P(i)}{Q(i)}\right). \tag{3.4}$$

  A smaller value indicates a closer match between the model and real data.

- **Mutual Information**: it quantifies the information obtained about one random variable by observing the other. It is defined as

$$I(G^{'}, G) = \sum_{g^{'} \in \mathcal{G}^{'}} \sum_{g \in \mathcal{G}} P(g^{'}, g) \log\left(\frac{P(g^{'}, g)}{P(g^{'})Q(g)}\right) \tag{3.5}$$

- **Normalised Mutual Information**, NMI: it adjusts the mutual information to account for the entropy of the variables, providing a normalized measure between 0 and 1:

$$NMI\left(G^{'}, G\right) = \frac{I\left(G^{'}, G\right)}{f\left(H\left(G^{'}\right), H\left(G\right)\right)}, \tag{3.6}$$

  where $f(\cdot)$ is a function of $H\left(G^{'}\right)$ and $H(G)$ such as square root, maximum value, minimum value, and sum.

- **Adjusted Mutual Information**, AMI: it corrects the mutual information for chance, offering a more accurate comparison. It is calculated as

$$AMI\left(G^{'}, G\right) = \frac{I\left(G^{'}, G\right) - \mathbb{E}\left[I\left(G^{'}, G\right)\right]}{\max\left(H\left(G^{'}\right), H\left(G\right)\right) - \mathbb{E}\left[I(G^{'}, G)\right]}, \tag{3.7}$$

  where $\mathbb{E}\left[I\left(G^{'}, G\right)\right]$ is the expected Mutual Information between $G^{'}$ and G under random distribution.

- **Pearson Correlation Coefficient**, ρ: it measures linear correlation between two sets of data. It is calculated as

$$\rho\left(G^{'}, G\right) = \frac{cov\left(G^{'}, G\right)}{\sigma\left(G^{'}\right)\sigma\left(G\right)}, \tag{3.8}$$

3.3 Experimental Validation

The experimental setup involved treating breast cancer cell with specific inhibitors of Argonaute 2 (AGO2), followed by comprehensive gene expression

analysis. The data for this analysis were extracted using microarray technology, which allows for the simultaneous examination of the expression levels of thousands of genes.

We evaluated three distinct sampling approaches to assess their effectiveness in molecular binding: Fast Sampling (FS), which considers only the score as extraction probabilities; Product Sampling, PS, which bases extraction on the product of molecules' expression and score; Competitive Model, CM, which uses a two-step extraction process, first based on expression and then on score.

Each model was applied to four different urn compositions, including balanced and unbalanced scenarios. In particular, we considered $\mathcal{S}_5^5$ and $\mathcal{S}_7^7$, i.e. scenarios made of urns composed of $100,000$ and $10,000,000$ molecules for each type, and $\mathcal{S}_5^7$ and $\mathcal{S}_7^5$, i.e. unbalanced scenarios with urns composed of $100,000$ genes and $10,000,000$ miRNAs and vice versa.

Furthermore, we tested each model under four different assumptions regarding the score. The first and second assumptions use scores directly from the chosen tool, but in the second, we shuffle these scores. The third and fourth assumptions retain a fraction of the miRNA-gene interactions, specifically those that are experimentally validated pairs obtained from the miRTarBase database [1]. The scores utilised in this study were PITA, MirandaScore, and MirandaEnergy, evaluated in sum, mean, and max configurations.

FS exhibited the lowest performance overall with correlation coefficients proximal to zero. The density overlap between the immunoprecipitated of the *in vitro* experiment and those of the model was minimal, suggesting that relying solely on the score for extraction probabilities is ineffective. Contrarily, PS showed better performance, with a greater density overlap observed when incorporating the expression of molecules. Table 3.1 shows the correlation coefficient obtained using the PS model [1]. The Shuffling and Validated columns represent the four assumptions, with YES indicating that shuffling and filtering of the scores were applied, respectively. CM showed no significant improvement, suggesting that the two-step extraction process does not provide additional benefits over simpler methods

Among the various score configurations and assumptions tested, the PITA score in the sum configuration consistently yielded the best results.

The evaluation across different urn compositions and score assumptions

[1]We have reported only the correlation for simplicity, as the correlation between the correlation and other metrics is close to one.

Table 3.1: Correlation coefficient between the immunoprecipitated from *in vitro* experiment and the output of the Product Sampling for different scores and urn compositions

Score	Shuffling	Validated	$\mathcal{S}_5^5$	$\mathcal{S}_7^5$	$\mathcal{S}_5^7$	$\mathcal{S}_7^7$
Max MirandaEnergy	NO	NO	0.40	0.40	0.40	0.40
	NO	YES	0.40	0.40	0.40	0.40
	YES	NO	0.39	0.38	0.40	0.38
	YES	YES	0.38	0.40	0.38	0.40
Max MirandaScore	NO	NO	0.39	0.38	0.38	0.38
	NO	YES	0.38	0.38	0.38	0.38
	YES	NO	0.39	0.39	0.39	0.39
	YES	YES	0.39	0.39	0.38	0.39
Max PITA	NO	NO	0.41	0.41	0.41	0.41
	NO	YES	0.41	0.41	0.41	0.41
	YES	NO	0.39	0.40	0.39	0.35
	YES	YES	0.38	0.38	0.40	0.38
mean MirandaEnergy	NO	NO	0.40	0.40	0.41	0.40
	NO	YES	0.40	0.40	0.41	0.40
	YES	NO	0.38	0.37	0.37	0.39
	YES	YES	0.39	0.38	0.39	0.39
mean MirandaScore	NO	NO	0.38	0.38	0.38	0.38
	NO	YES	0.38	0.38	0.39	0.38
	YES	NO	0.39	0.39	0.39	0.39
	YES	YES	0.39	0.38	0.39	0.39
mean PITA	NO	NO	0.41	0.41	0.41	0.41
	NO	YES	0.41	0.41	0.40	0.41
	YES	NO	0.40	0.39	0.39	0.38
	YES	YES	0.35	0.39	0.39	0.40
sum MirandaEnergy	NO	NO	0.38	0.39	0.38	0.39
	NO	YES	0.39	0.38	0.38	0.39
	YES	NO	0.36	0.38	0.38	0.38
	YES	YES	0.39	0.38	0.38	0.40
sum MirandaScore	NO	NO	0.38	0.37	0.38	0.37
	NO	YES	0.37	0.37	0.37	0.37
	YES	NO	0.39	0.38	0.34	0.39
	YES	YES	0.34	0.37	0.37	0.39
sum PITA	NO	NO	0.39	0.39	0.39	0.39
	NO	YES	0.40	0.39	0.40	0.40
	YES	NO	0.38	0.36	0.38	0.36
	YES	YES	0.39	0.38	0.38	0.37

highlights the importance of integrating molecular expression and score metrics to improve binding evaluation. However, the lack of improvement in the competitive model suggests that simpler, more direct methods may be preferable in certain contexts.

3.4 Discussion

The study under investigation concerns the role of miRNA regulation in breast cancer cells, utilizing microarray data to extract and analyze relevant information. AGO2 is a critical component of the RISC and plays a significant role in regulating gene expression through microRNA based target recognition.

The experimental setup consisted into the separation of RNA binded to microRNAs through the immunoprecipitation of AGO2, followed by the gene expression analysis of three kind of samples: IN, IP and FT. The data for this analysis were extracted using microarray technology, which allows for the simultaneous examination of the expression levels of thousands of genes.

The evaluation of miRNA-gene interactions is crucial for understanding gene regulation mechanisms. Our study aimed to determine whether the binding score is a significant predictor of miRNA-gene interactions. To achieve this, we developed and tested three distinct models: a fast sampling model (FS), a product sampling model (PS), and a competitive model (CM). Each model incorporates different approaches to integrate the binding score and gene expression data to assess their predictive accuracy. In particular, FS considers only the binding score, providing a straightforward and rapid assessment of its predictive capability. PS considers the product of the binding score and the molecules' expression levels, aiming to enhance the prediction by incorporating expression data. CM involves two extraction steps, the former based on gene and miRNA expressions and the latter based on the score, creating a more complex framework to evaluate the binding interactions.

To evaluate the predictive accuracy of these models, we compared their outputs with experimental immunoprecipitation data, which serves as the benchmark for actual binding events. The accuracy of the models was measured using several metrics.

Our findings indicate that FS consistently produced poor results across all metrics, highlighting the importance of considering gene and miRNA expression levels in the prediction process, as ignoring these factors diminishes

the model accuracy. In contrast, the product sampling model, which incorporates the binding score and molecules expression through their product, demonstrated superior performance. While more intricate, the competitive model did not outperform the product sampling model, suggesting that the additional complexity did not translate into better predictive accuracy.

We have utilised several scores in their different configuration. These scores are derived from a consistent list of miRNA-gene pairs, ensuring that the fundamental data set remains unchanged across comparisons. Importantly, the role of these scores is not to predict which pairs will interact but to provide an estimate of the affinity between miRNAs and their target genes. This affinity is then interpreted as the probability of a binding event occurring.

Among the various scores available, SumPITA appears to be the most effective tool, although the other tools have also demonstrated valuable results, demonstrating behaviour that aligns closely with our expectations for accurate prediction models. PITA distinguishes itself by utilising extensive binding information, specifically the binding energy between miRNA and the target gene and the energy required to facilitate this binding. The latter includes the energy necessary to destabilise the secondary structure of RNA, which might otherwise impede the binding site. Hence, PITA provides a more accurate reflection of the true binding dynamics, a feature that translates into the enhanced performance observed with SumPITA.

Our empirical evidence supports the superiority of SumPITA. It outperforms other methods in comparative analyses. The rationale is straightforward: multiple binding sites increase the likelihood of binding events, an aspect that SumPITA effectively captures. This characteristic is not adequately reflected in alternative scoring methods or other PITA evaluations, such as Mean or Max, which tend to average out or oversimplify the binding potential.

3.5 Conclusion

In conclusion, our investigation into the role of AGO2 inhibition in breast cancer cells, utilizing microarray data, has provided significant insights into miRNA-gene interaction prediction. Through an experimental setup involving AGO2 inhibitors and subsequent gene expression analysis, we employed three distinct models - fast sampling (FS), product sampling (PS), and competitive model (CM) - to evaluate the predictive accuracy of binding scores.

Our results underscore the importance of incorporating gene and miRNA expression levels, as the FS model, which neglected these factors, consistently underperformed. Conversely, the PS model, particularly when utilizing the PITA scores, demonstrated superior performance, highlighting the efficacy of integrating binding scores with expression data. The CM, despite its complexity, did not surpass the PS model predictive accuracy. Our findings affirm the superiority of SumPITA, which effectively captures the increased likelihood of binding events associated with multiple binding sites, thereby providing a more accurate reflection of true binding dynamics.

Future Researches

This section focuses on current and future researches that are strictly related to the present thesis' analyses.

3.6 Upgrade of LAA detection algorithm to extract more LAA features

In Chapter 1, we propose a new method to separate LAA from LA without requiring human intervention. By product, the method also extracts several features that characterise the LAA. An improvement of the current state of the algorithm may concern the computation of several other features. For instance, Fractal Analysis has merged as a novel approach to explore the geometrical characteristics and self-similarity of the LAA across different scales. Indeed, the fractal nature of LAA, which is articulated in its trabeculae structure, has been largely unexplored. Furthermore, while Fractal Analysis has been extensively applied in the biomedical field, such as in investigating tumor growth patterns, endocardial trabeculation, and pulmonary branching, there are very few applications for the LAA.

Algebraic fractals typically require thousands or millions of iterations before their fractal nature becomes apparent, and as a result, they are often visualised using computer software. It is no surprise that widespread appreciation for the complexity of fractals emerged only with the advent of computers in the 1980s, largely attributed to the pioneering work of Mandelbrot and Mandelbrot [95]. Fractal geometry can be viewed as a natural extension of traditional (Euclidean) geometry [133]. In traditional geometry, work is typically done with integer dimensions representing the length exponent, such as surface = length^2 or volume = length^3. The exponent itself defines the dimension. However, fractal geometry introduces a novel concept where measurements exhibit non-integer or fractional changes as the unit of measurement is altered. The gov-

erning parameter, FD, is known as the fractal dimension [133, 116, 143]. This index, comparing other types of dimensions (Lebesgue Dimension, Hausdorff Dimension), is a descriptive and quantitative metric used as an endeavour to approximate a singular numerical value for a characteristic (complexity) of an object using a sample of data obtained from the object [80].

Fractal Analysis has recently gained recognition in the medical field for its application in quantifying anatomical complexity. Granted that biological structures are not mathematically speaking "fractals" due to a high degree of self-similarity observed within selected anatomical structures, fractal analysis has already been accurately used to quantify irregularity within retinal vasculature, neuronal grey matter networks within the spinal cord, trabecular bone density, pulmonary branching and endocardial trabeculation in the presence of left ventricular non-compaction, tumour growth patterns. Fractal analysis has already accurately investigated the LV endocardium infrastructure. Lei et al. [86] describes Fractal Geometry as a tool to illustrate LAA Morphology in order to predict thrombosis and stroke in patients with AF.

To the best of our knowledge, all the algorithms developed to compute the Fractal Dimension of an object rely on 2D images. Our research focuses on a dimensionally independent algorithm. Thus, our approach is a potentially novel contribution in this domain. Based on the box-counting method, the algorithm has already correctly computed the Fractal Dimension of mathematical objects, such as the Sierpinski Triangle, Carpet, Tetrahedron and the Menger Sponge.

3.7 Improvement of the Clustering Framework

Chapter 2 delves into the intricacies of developing a robust clustering framework of the LAA morphologies. The foundation of this framework relies heavily on the LAA features but suffers from a limited number of patients, making the framework promising but still far from its definitive form. As the availability of patients continues to grow, their impact on the clustering framework becomes increasingly significant.

A larger dataset would allow the clustering algorithms to capture subtle patterns and variations that smaller datasets might miss. A key advantage of a larger dataset is the enhancement of the generalization capabilities of the clustering framework. More patients will likely include various examples, in-

cluding rare and outlier cases. This diversity ensures that the clustering model is exposed to a broad spectrum of scenarios, improving its ability to generalize to new, unseen data. Clustering algorithms often require tuning hyperparameters, and a larger dataset provides a more robust platform for hyperparameter optimization. Techniques such as cross-validation become more reliable as the larger volume of data ensures that each fold is representative of the entire dataset.

3.8 Binding Probabilities Optimisation

A future extension of this work could focus on optimizing the binding probabilities. Specifically, it would involve integrating and leveraging the three scores provided by PITA, Miranda, and TargetScan. The objective would be to maximize the overlap between the observed IP and the predicted one generated by the model. This optimization would be achieved by constraining the binding probabilities to closely align with the scores obtained from PITA, Miranda, and TargetScan. This approach aims to enhance the accuracy and reliability of the model by ensuring that the binding probabilities reflect the consensus of these established scoring systems.

Conclusions

The present thesis offers an innovative approach to analysing the LAA and evaluating molecular interaction. The system described in the first chapter effectively addresses critical gaps in the morphological analysis of the LAA. The novel method overcomes the uncertainty in distinguishing the auricula from the left atrium by automatically detecting a cutting plane and employing quantitative, reproducible procedures to compute LAA features. This approach successfully detected all appendages in a sample of 86 control patients without human intervention, promising enhanced comparability in future morphological studies of the LAA.

Furthermore, we have laid the groundwork for a novel approach to understanding LAA morphology through a feature-based clustering methodology. By embracing a data-driven approach and integrating various clustering methodologies, we have proposed a more objective and scientifically robust framework for delineating LAA identity groups. This approach, which considers a broad spectrum of morphological characteristics, revealed two distinct LAA groups characterised by size and internal complexity through K-means and Fuzzy K-means clustering. While these groups may not directly align with existing classifications in the literature, the characterisation based on multiple extracted LAA features offers significant insights. Future research should aim to expand the sample size to enhance the quality and robustness of our results, exploring the clinical implications of these identified LAA groups concerning clot-formation risk and treatment strategies. This study marks a significant advancement in LAA morphology classification, presenting a promising alternative for future research and clinical applications.

Additionally, we introduced the rewiring algorithm to evaluate miRNA interactions, comprising three phases: data pre-processing, the competitive interaction model, and the evaluation procedure. This algorithm contributes to a broader understanding of molecular interactions.

An overall view of the results presented in this thesis strongly supports improved LAA detection and characterisation of its morphology. In particular, the fully automated method allows researchers to make comparisons across future studies. Moreover, the feature-based LAA groups lead to a more resonable grouping of the LAA and overcome subjectivity in classifying LAA morphologies. Finally, we have laid the groundwork for a more complete evaluation of the miRNA-mRNA interaction through a competitive model.

List of Figures

1.1 (a) Anatomical position of the Left Atrial Appendage. (b) Visual representation of two distinct LAA cases, highlighting the anatomical features such as the ostium, neck, primary and secondary lobes, and trabeculae. 13

1.2 The first row of the figure shows the automatic detection developed by Masci et al. [97], where the shape is coloured using the SDF (a), and the appendage detected (b). (c) shows all the data utilised in their work. (d) shows the erosion-dilation detection algorithm of Lei et al. [86]. 18

1.3 Example of branches and twigs recognition in the analised data. 20

1.4 Flow chart of the left atrial appendage detection method. 21

1.5 Given a triangulated surface mesh (a), two skeletons are extracted: a simple one (b) and a complex one (c). The facet-skeleton distances are computed using the simple skeleton, and the distances are used to map the mesh surface (d). The shape is divided into two groups (red and green) using a threshold measure (e). The borders of the groups are smoothed by evaluating the membership affinity of the faces (f). The LAA (cyan) is detected by identifying the connected component of the red group with the most complex inward skeleton, and a plane is used to cut it with a net cut (g). The LAA's skeleton (cyan) is obtained by cutting the complex skeleton with the plane (h). . . 24

1.6 Affinity refinement used to fill little irregularities. The candidate facet f_0 belonging to a given group changes its membership (cases a and b) if it has more than 50% of its neighbours belonging to the other group. 27

1.7 (a) and (b) A cut along a trabecula generates multiple holes. (c) The cutting plane identifies a shape containing the PVs with a ellipse goodness of fit $G = 3,700$. 29

1.8 (a) The red path is used to compute the length of the ostium perimeter. The green faces are used to compute the area of the ostium. (b) The ellipse fitted to compute the eccentricity of the ostium. (c) The skeleton of the appendage. (d) The red path highlights the trabeculae (e) by showing how skeletal loops help to identify them. (f) The blue path highlight the main path. (g) The orange path shows the shortest path that preserves at least 95% of the length of the main path. (h) The length of the blue and black paths (arch-cord) are used to compute the tortuosity, while θ and π are used to compute the curvature. 30

1.9 Box-plot of the features extracted by our method. Unit of measurement and domain of each feature is reported in Table 1.1. . 34

2.1 Correlation plot among features extracted by the method before (a) and after (b) the highly correlated features removal 73

2.2 Number of clusters identified in K-means approach by elbow (top), gap statistic (middle) and silhouette (bottom) methods. . 74

2.3 Number of clusters identified in Fuzzy K-means approach by Fuzzy Silhouette (top left), Modified Partition Coefficient (top right), Partition Coefficient (middle left), Partition Entropy (middle right), and Silhouette (bottom left). Colours shows results for three fuzzy parameters: $m = 1.5$ (red), $m = 2.0$ (green), and $m = 2.5$ (cyan) 75

2.4 Dendrograms of the hierarchical clustering approaches: (a) agglomerative with average linkage; (b) agglomerative with McQuitty linkage; (c) agglomerative with complete linkage; (d) divisive Diana. Horizontal lines represent the dendrogram cut at a $d = \sqrt{2}$, corresponding to a null correlation between units. . 76

2.5 BIC values for combination of number of clusters and within-group covariance matrix parametrisation. Best three models are identified through red, green and blue points. 77

2.6 Networks and communities detected through (a) Edge Betweenness in a Planar Maximally Filtered Graph configuration, (b) Louvain and (c) Leading Eigen methods using a Leave-One-Out robustness metric based on positive correlation matrix. 79

2.7 Graphical representation of the Macro F1 score (see Figure 2.7a), Weighted F1 (see Figure 2.7b), and average Balanced Accuracy (see Figure 2.7c) obtained during the K-fold procedure to evaluate the reliability of partitions. The partitions are depicted on the y-axis, while the x-axis illustrates the supervised classification algorithm. A higher value of the indices indicates superior performance in classifying new units. Clustering techniques with a null value in either Macro F1 score or Weighted F1 were omitted to enhance graphical representability. 87

2.8 Similarity measures for partition comparison. (a) Precision, (b) Recall, (c) F1 score, (d) Jaccard index, (e) Mutual Information, (f) Normalised Mutual Information, (g) Adjusted Mutual Information. 88

2.9 Boxplot by group of the features. 89

3.1 Scenario space. Gray cells represent scenarios that do not make sense to analyse because the total expression is too small. Blue cells refer to balanced scenarios, where the cardinality of gene and miRNA urns are equal. Green (orange) cells refer to unbalanced scenarios where the number of miRNA (gene) is greater than the gene (miRNA) one. 103

List of Tables

1.1 Extracted features of the LAA. 33

1.2 Features extracted by the main literature methods are denoted by the 💻 symbol if their computation is automated, and by ✋ if they are manually derived. Features that are not extracted at all remain unmarked. 36

2.1 Clustering approaches utilised to find LAAs partition. 55

2.2 Parameterisations of the within-group covariance matrix $\mathbf{\Sigma}_k$ and the corresponding geometric characteristics. Source Scrucca et al. [129]. 62

2.3 Removed variable by correlation evaluation 73

2.4 Number of clusters for each partition found. Cluster cardinalities are in descending order, and no correspondences belong between different methodologies. The cluster cardinality column shows at most six clusters, if present. (n) indicates the presence of n groups of cardinality equal to 1. $(*)$ indicates the presence of just one cluster with the entire data sample belonging to it. . 86

2.5 Main statistics of the features by group. 90

3.1 Correlation coefficient between the immunoprecipitated from *in vitro* experiment and the output of the Product Sampling for different scores and urn compositions 107

A.1 Detection results of the system discussed in Chapter 1. i

B.1 Optimal partition of the sample data into two groups. xv

Acronyms

AF: Atrial Fibrillation

AGO2: Argonaute 2

ALCA: Average Linkage Cluster Analysis

ANN: Artificial Neural Networks

BA: Balanced Accuracy

BIC: Bayesan Information Criterion

CCTA: Coronary Computed Tomography Angiography

CLCA: Complete Linkage Cluster Analysis

CM: Competitive Model

CN: Complete Network

CRF: Conditional Random Field

CT: Computed Tomography

CTp: Planar CT

CTsg: Segmented CT

EM: Expectation-Maximization

ESC: European Society of Cardiology

FCN: Fully Convolutional Network

FS: Fast Sampling

FT: Flow-Through Solution

GMM: Gassian Mixture Model

GN: Group Normalization

IN: Input or Starting Solution

IP: Imhuno Precipitated

KNN: K-Nearest Neighbours

LA: Left Atrium

LAA: Left Atrial Appendage

LV: Left Ventricle

MAP: Maximum a Posteriori Decision

MCC: Matthews Correlation Coefficient

MDCT: Multidetector CT

MFC: Mean Curvature Flow

MLR: Multinomial Logistic Regression

MRA: Magnetic Resonance Angiografy

MRI: Magnetic Resonance Imaging

MST: Minimum Spanning Tree

NB: Naïve Bayes

NICE: National Institute for Health and Care Excellence

PCA: Principal Component Analysis

PMFG: Planar Maximally Filtered Graph

PS: Product Sampling

PV: Pulmonary Vein

RF: Random Forest

RNA: Ribonucleic Acid

ROC-AUC: Area Under the Curve Operating Characteristic

ROI: Region of Interest

RPROP: Resilient Propagation Algorithmi

ReLU: Rectified Linear Unit

SDF: Shape Diameter Function

SLCA: Single linkage cluster analysis

SVM: Support Vector Machine

TEE: Transesophageal Echocardiography

WSS: Within-Cluster Sum of Square

mRNA: Messenger RNA

miRNA: Micro RNA

Author contributions

Different authors contributed importantly to the projects on which the thesis is based. The following items highlight the main contributions:

- Chapter 1: Vincenzo Martorana[2] developed the method related to LAA detection and feature extraction and wrote the chapter supervised by Michele Tumminello[1], Giorgia Bosi[3], Gaetano Burriesci[2,4], and Claudia Coronnello[3]. Matthew Lee[2] and Rafizul Islam MD[2] have segmented all the data patients and provided the 3D data used as the model input.

- Chapter 2: Vincenzo Martorana[1] performed statistical data analysis and wrote the chapter supervised by Michele Tumminello.

- Chapter 3: Vincenzo Martorana[1] performed statistical data analysis and wrote the chapter supervised by Claudia Coronnello[2] and Michele Tumminello[1]. Giorgio Bertolazzi[1] and Vincenzo Martorana[1] worte the algorithm. Nicolina Sciaraffa[2] performed the RIP experiments and the microarrays.

[2]Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy

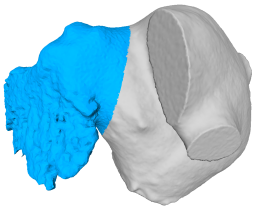
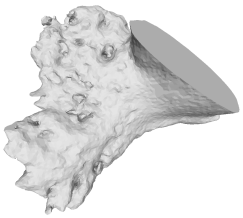
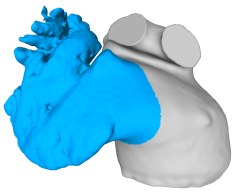
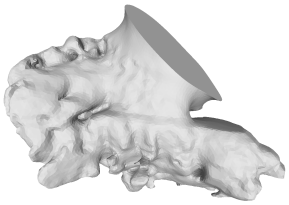
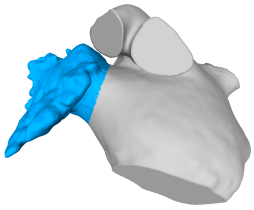
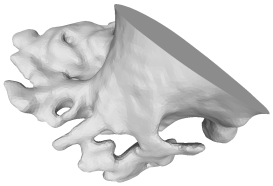
[3]Mechanical Engineering Department, University College of London, London, United Kingdom

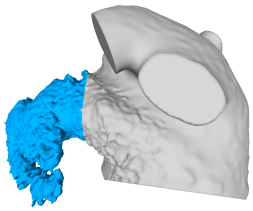
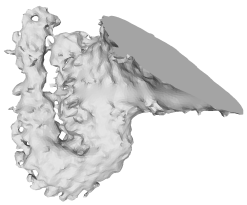
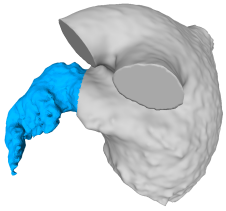
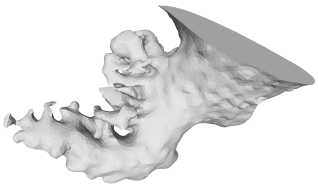
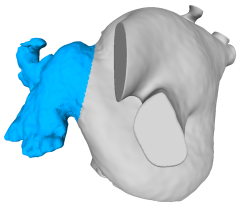
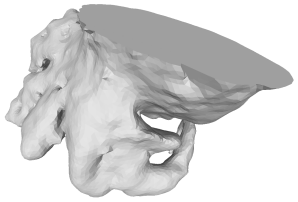
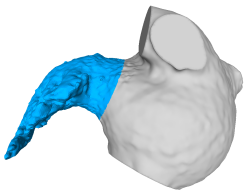
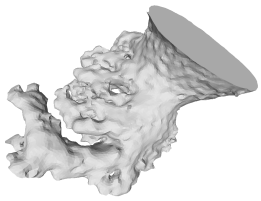
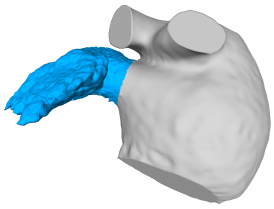
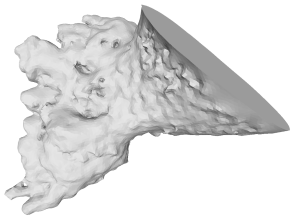
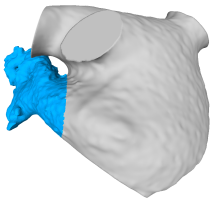
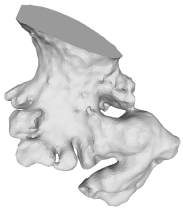
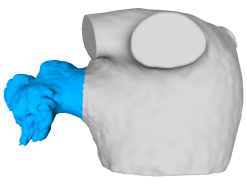
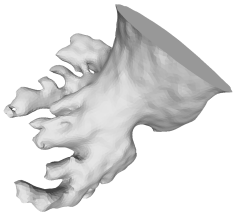
[4]Fondazione Ri.MED, Palermo, Italy

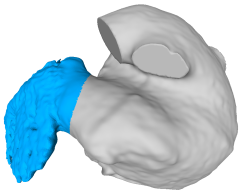
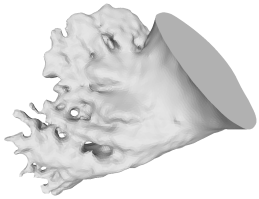
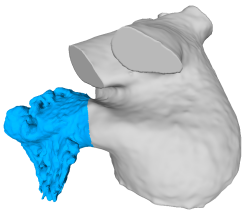
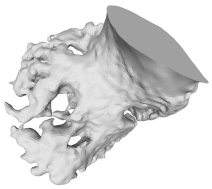
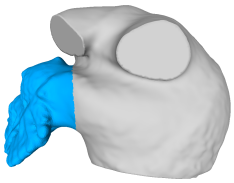
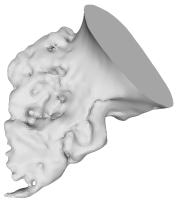
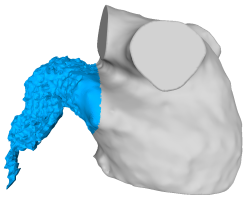
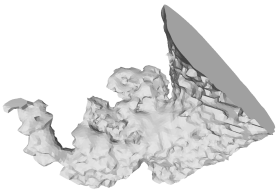
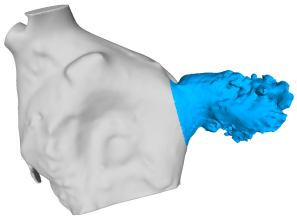
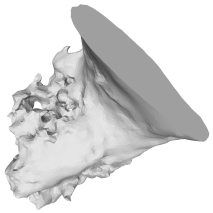
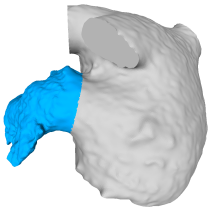
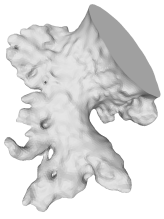
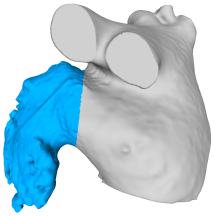
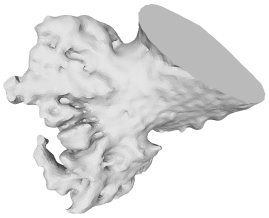
Appendix A

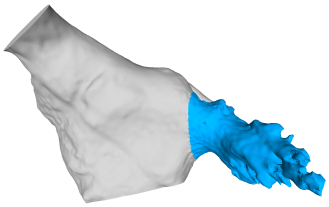
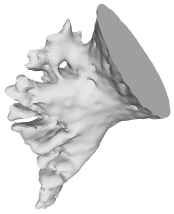
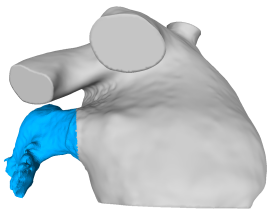
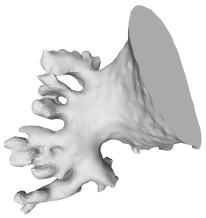
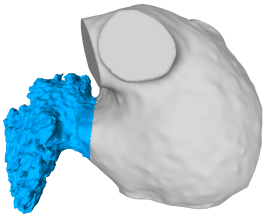
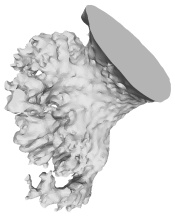
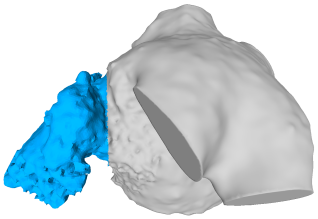
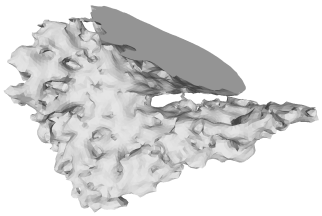
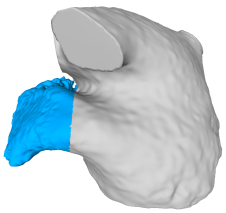
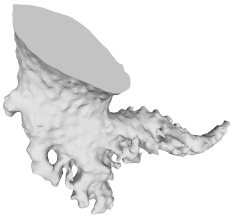
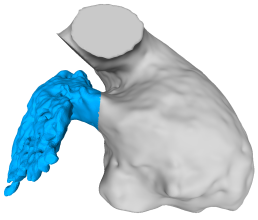
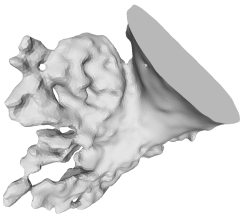
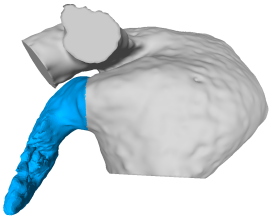
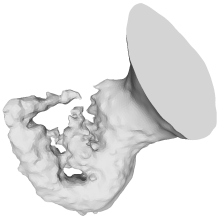
The following table shows the results of the system discussed in Chapter 1. Particularly, the first column shows the anonymised patient ID. In the second column, the table shows the LA+LAA shape. The system uses the identified plane to separate the LAA (blue) from the LA (grey). Some meshes are shown from the opposite view because PVs cover the LAA in the front view. In the last column, we present the LAAs in a view to show the ostium and as much of the LAA shape as possible.

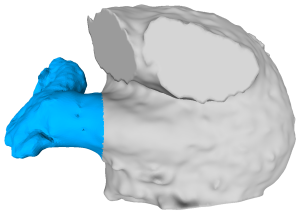
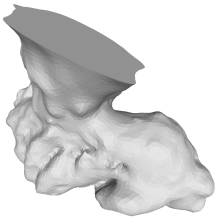
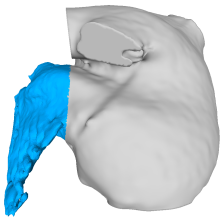
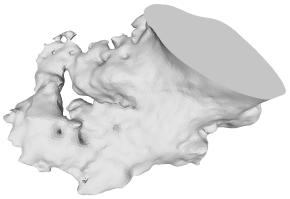
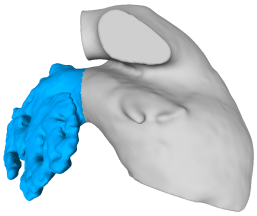
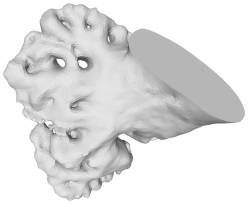
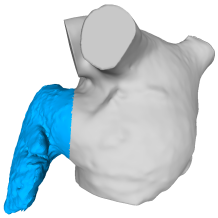
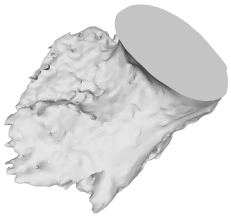
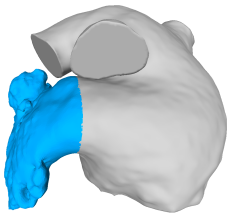
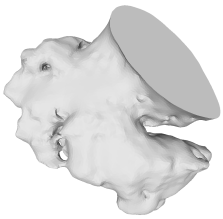
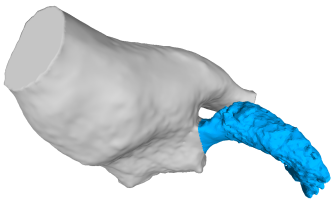
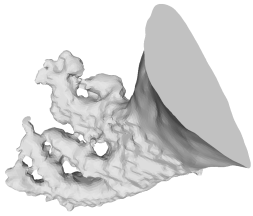
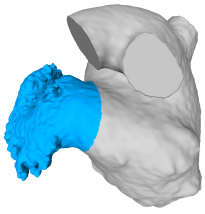
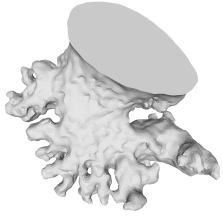
Table A.1: Detection results of the system discussed in Chapter 1.

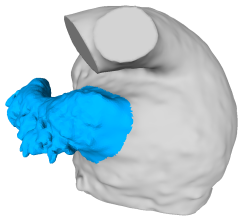
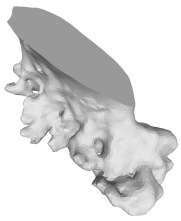
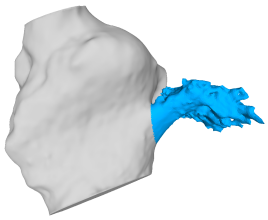
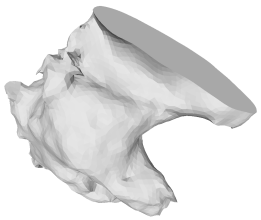
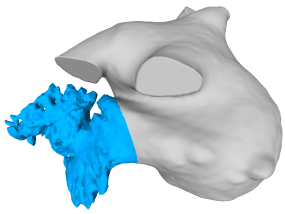
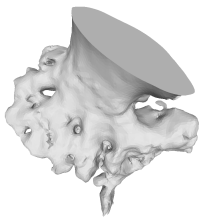
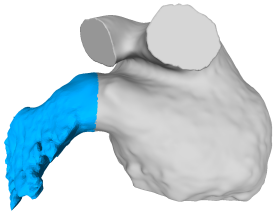
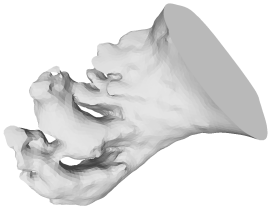
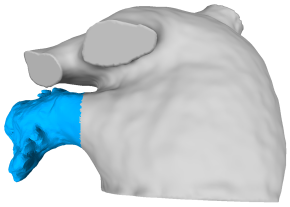
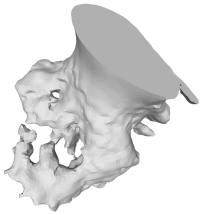
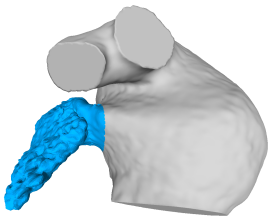
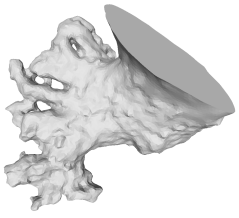
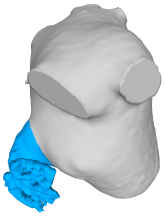
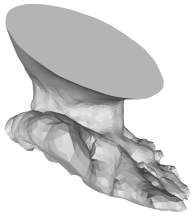
Patient	**LA + LAA + Cutting Plane**	**LAA**
1BA		
2AA		
4AI		

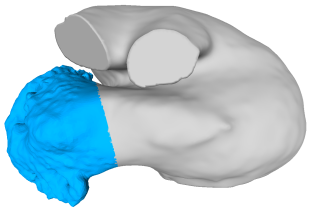
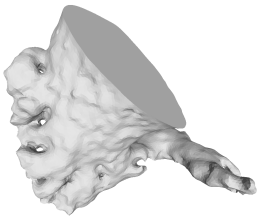
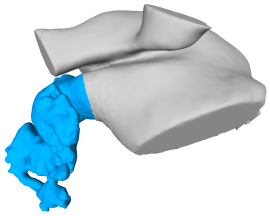
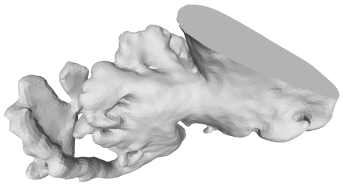
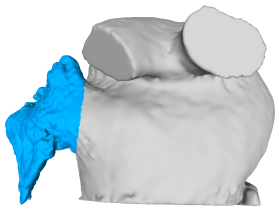
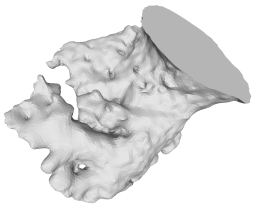
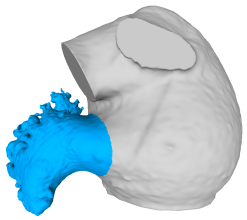
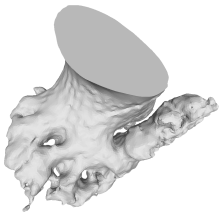
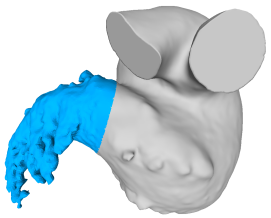
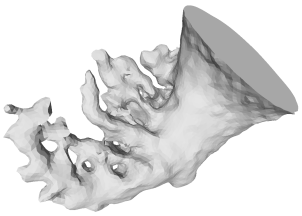
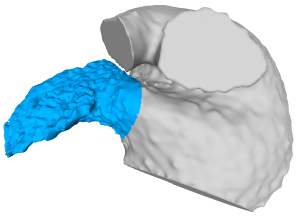
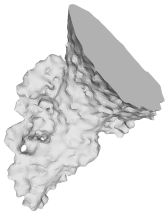
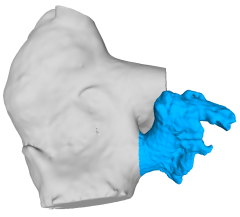
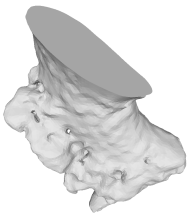
Patient	LA + LAA + Cutting Plane	LAA
5ML		
6HD		
7ME		
8BM		
9JJ		
10PA		
11CA		

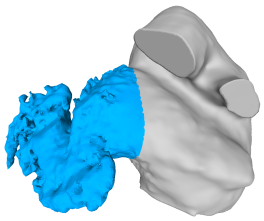
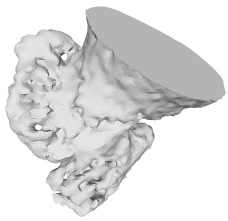
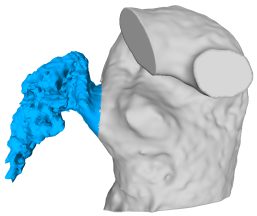
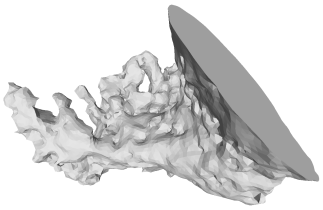
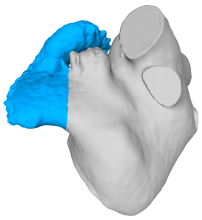
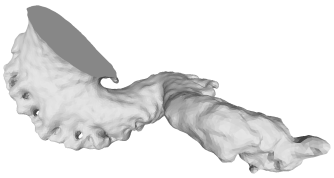
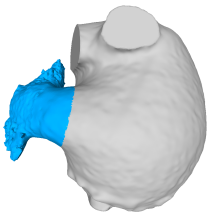
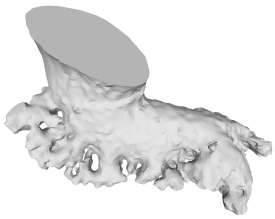
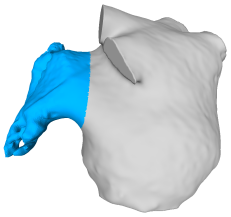
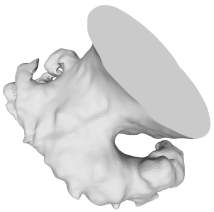
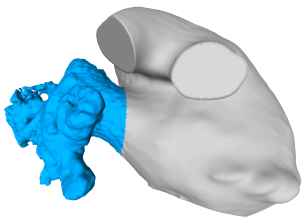
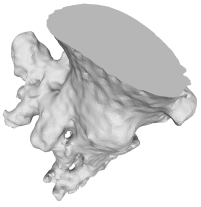
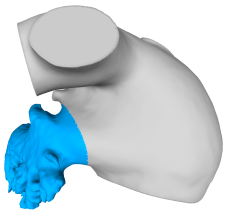
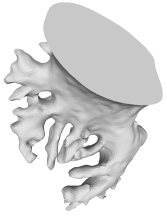
Patient	LA + LAA + Cutting Plane	LAA
12LA		
13MM		
15ML		
16NC		
17HM		
18SA		
19GB		

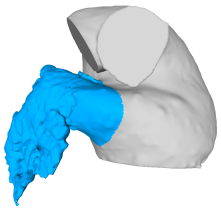
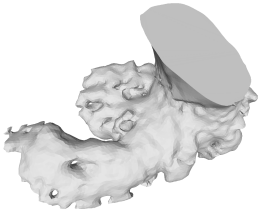
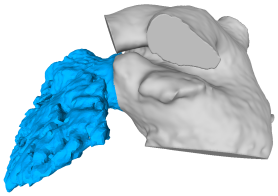
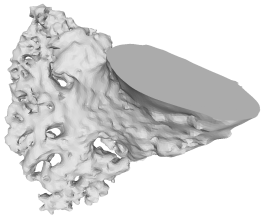
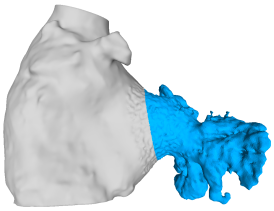
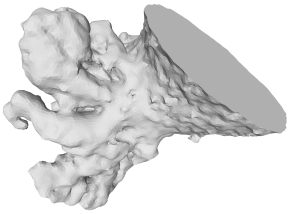
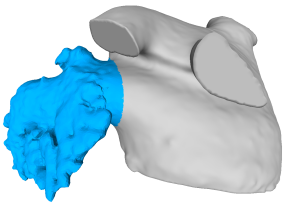
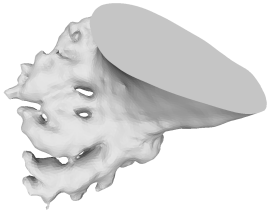
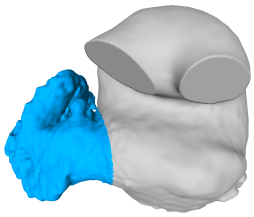
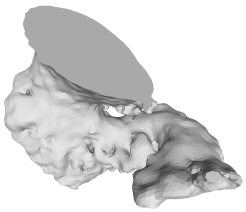
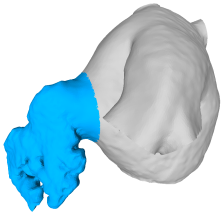
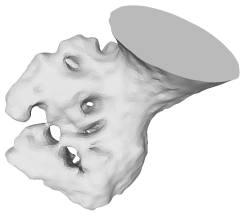
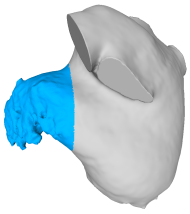
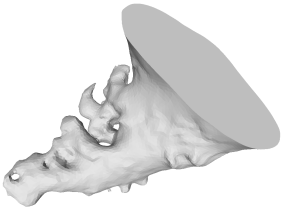
Patient	LA + LAA + Cutting Plane	LAA
24AM		
27GA		
28AG		
29AS		
30SM		
31PC		
33YE		

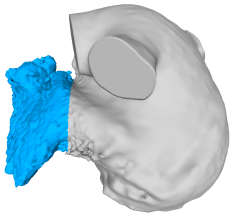
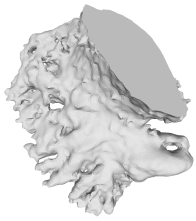
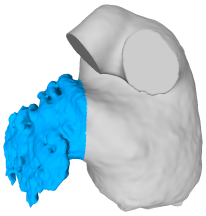
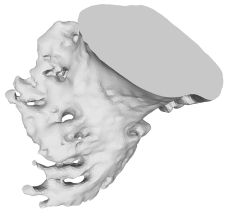
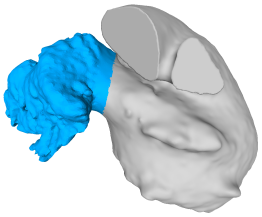
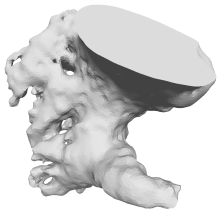
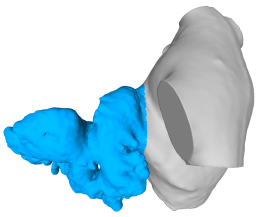
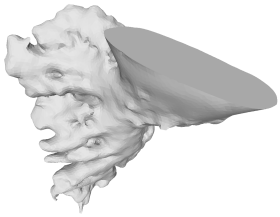
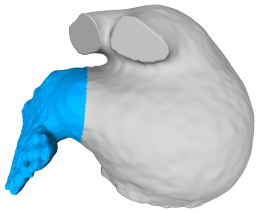
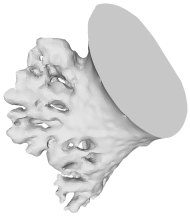
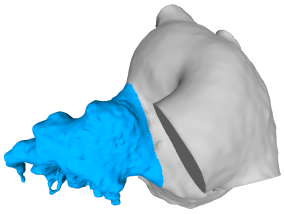
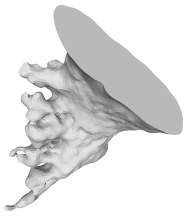
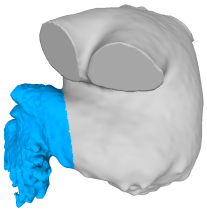
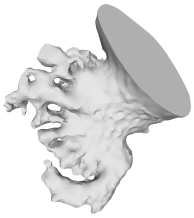
Patient	LA + LAA + Cutting Plane	LAA
34RS		
35AE		
36MJ		
38BJ		
39MF		
40ET		
41MR		

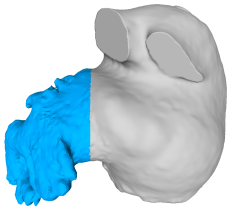
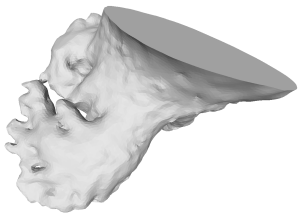
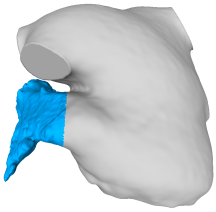
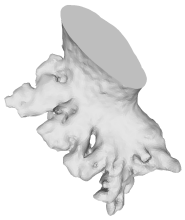
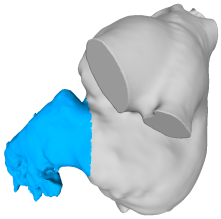
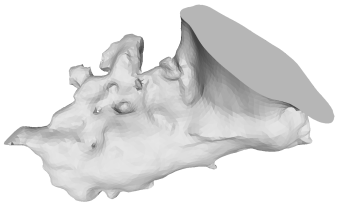
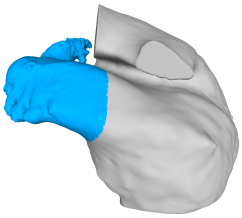
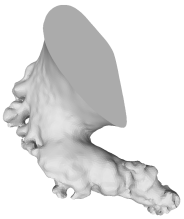
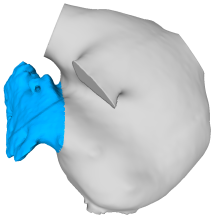
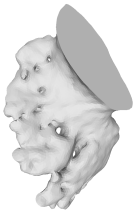
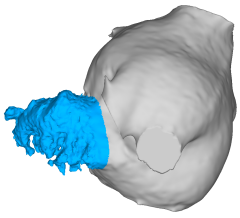
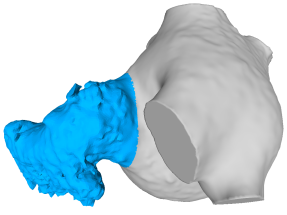
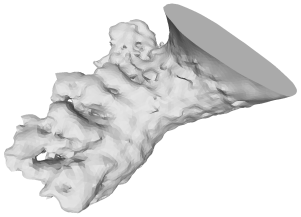
Patient	LA + LAA + Cutting Plane	LAA
42BK		
44BM		
45DC		
46RR		
48PA		
50MA		
51MJ		

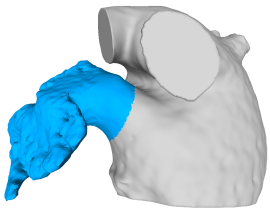
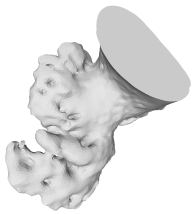
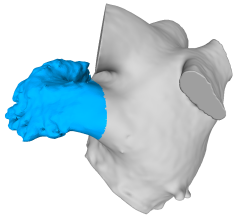
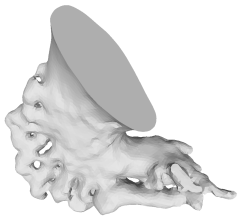
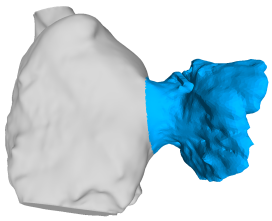
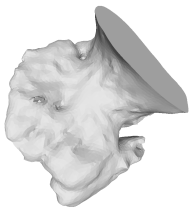
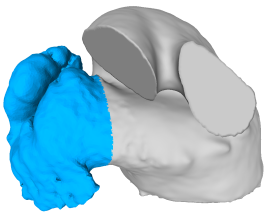
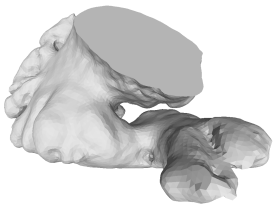
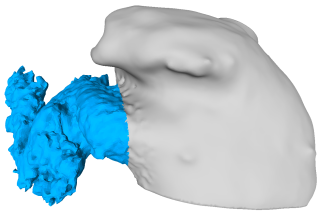
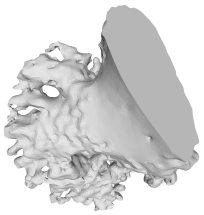
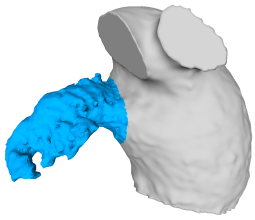
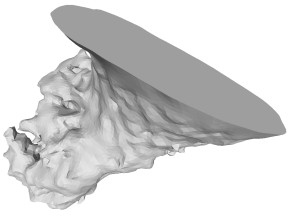
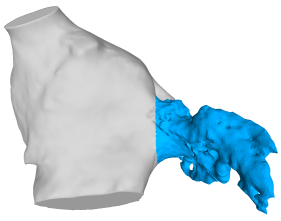
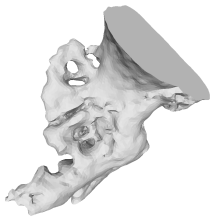
Patient	LA + LAA + Cutting Plane	LAA
53SD		
54SD		
56JP		
57SL		
58BH		
61MA		
62CH		

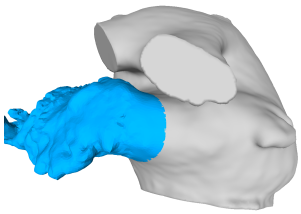
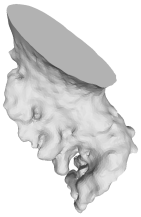
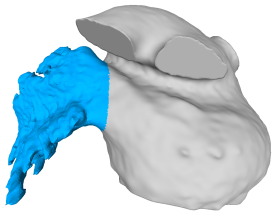
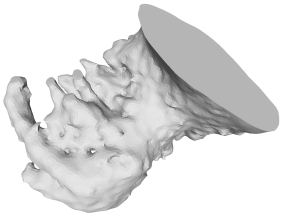
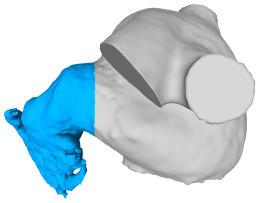
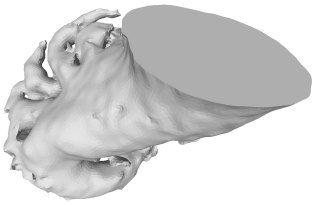
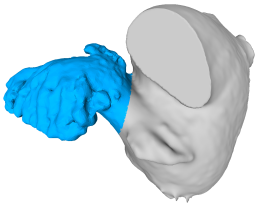
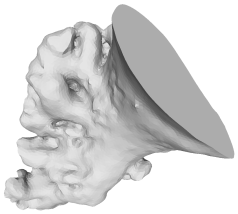
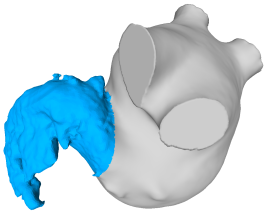
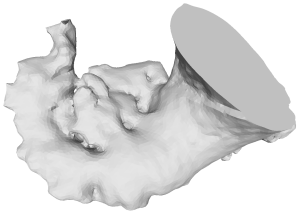
Patient	LA + LAA + Cutting Plane	LAA
65PE		
67RA		
68CF		
69AR		
70GH		
71HB		
72VS		

Patient	LA + LAA + Cutting Plane	LAA
73HM		
74AO		
75ST		
77MC		
78SE		
80AL		
81CA		

Patient	LA + LAA + Cutting Plane	LAA
82DR		
83BA		
84MP		
85GP		
86EJ		
87HS		
88RM		

Patient	LA + LAA + Cutting Plane	LAA
89IH		
90NJ		
91PF		
92MJ		
93TL		
94SJ		
95WJ		

Patient	LA + LAA + Cutting Plane	LAA
96DP		
98AA		
100DS		
101GM		
102EJ		
103HT		
104BR		

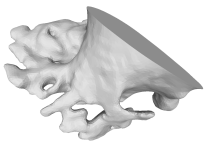
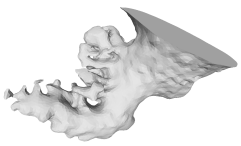
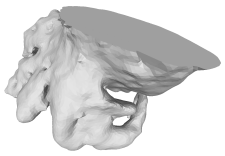
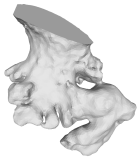
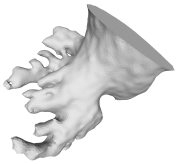
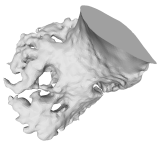
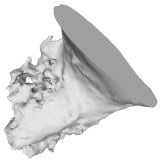
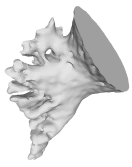
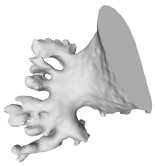
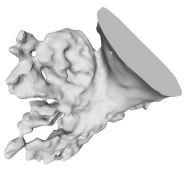
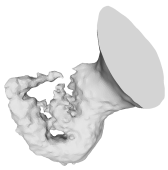
Patient	LA + LAA + Cutting Plane	LAA
105MJ		
106WC		
108MK		
110BC		
111DS		

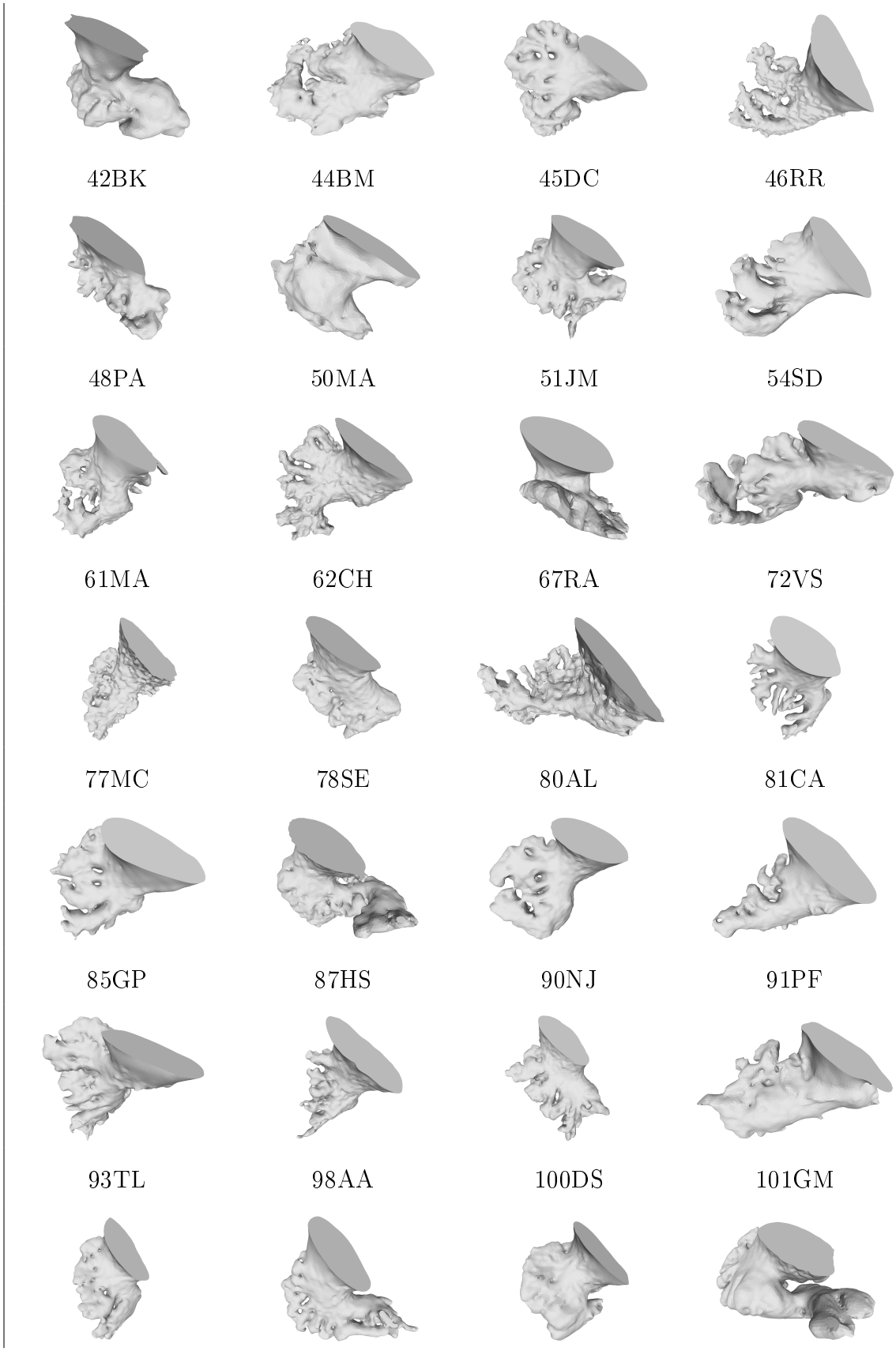
Appendix B

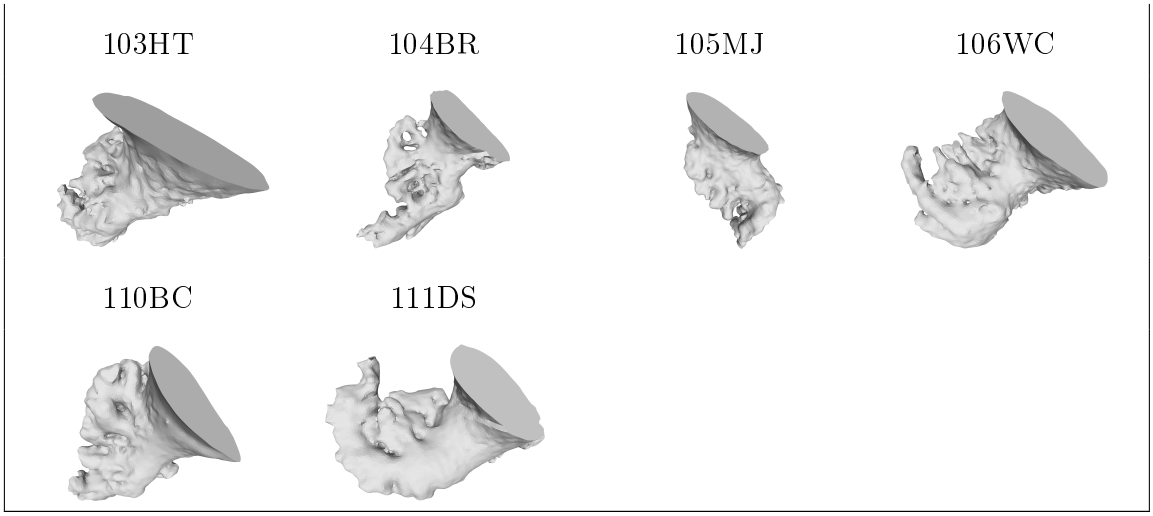
In the following table, we show the two groups that represent the best partition of the sample data. Group A is composed of LAAs with a smaller volume than the one in Group B. Length of the main path, ostium area, and perimeter are lower in Group A than in Group B. LAAs in Group B present more trabeculae and greater complexity of the main path than the ones in Group A.

Table B.1: Optimal partition of the sample data into two groups.

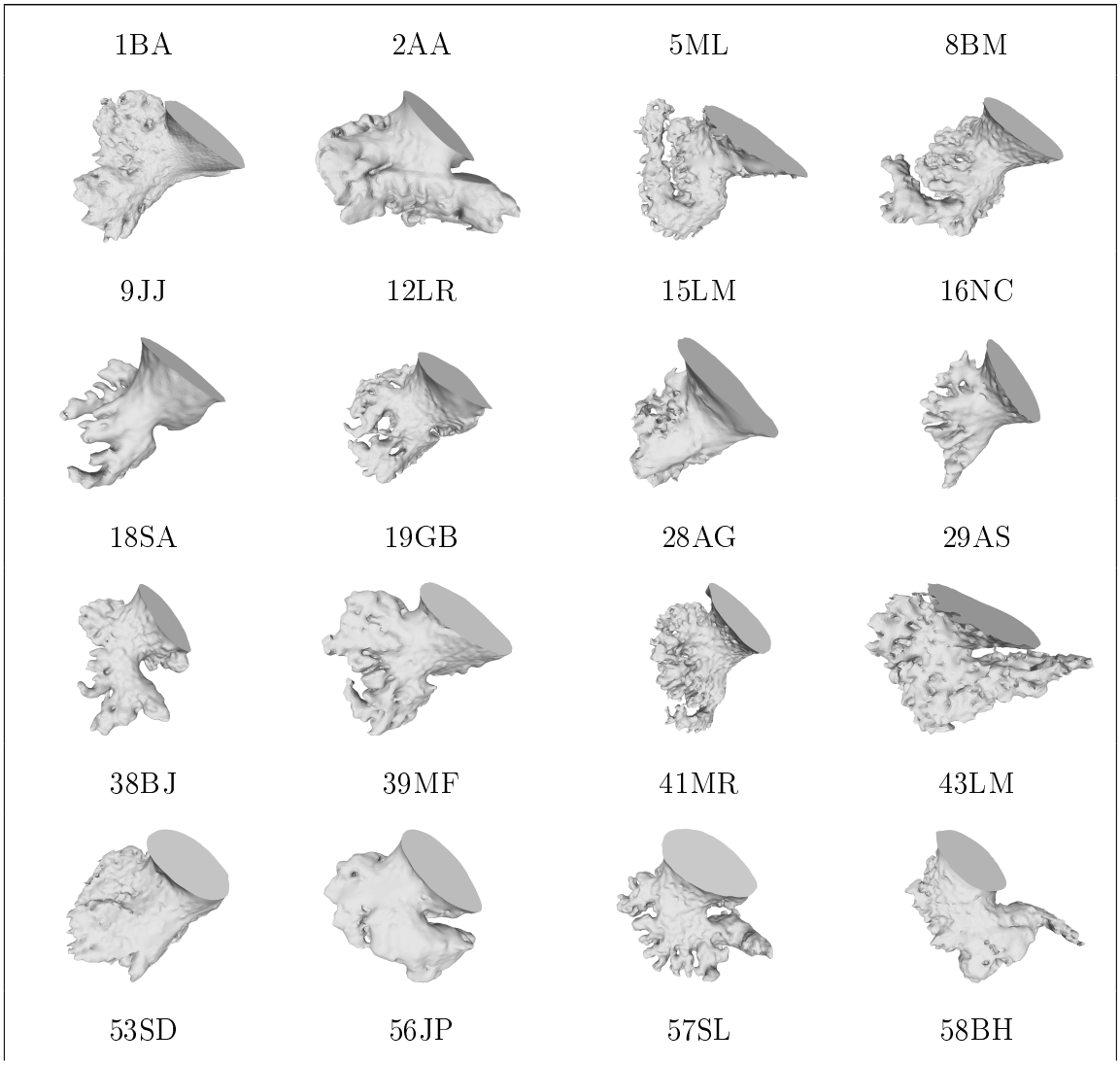
Cluster A

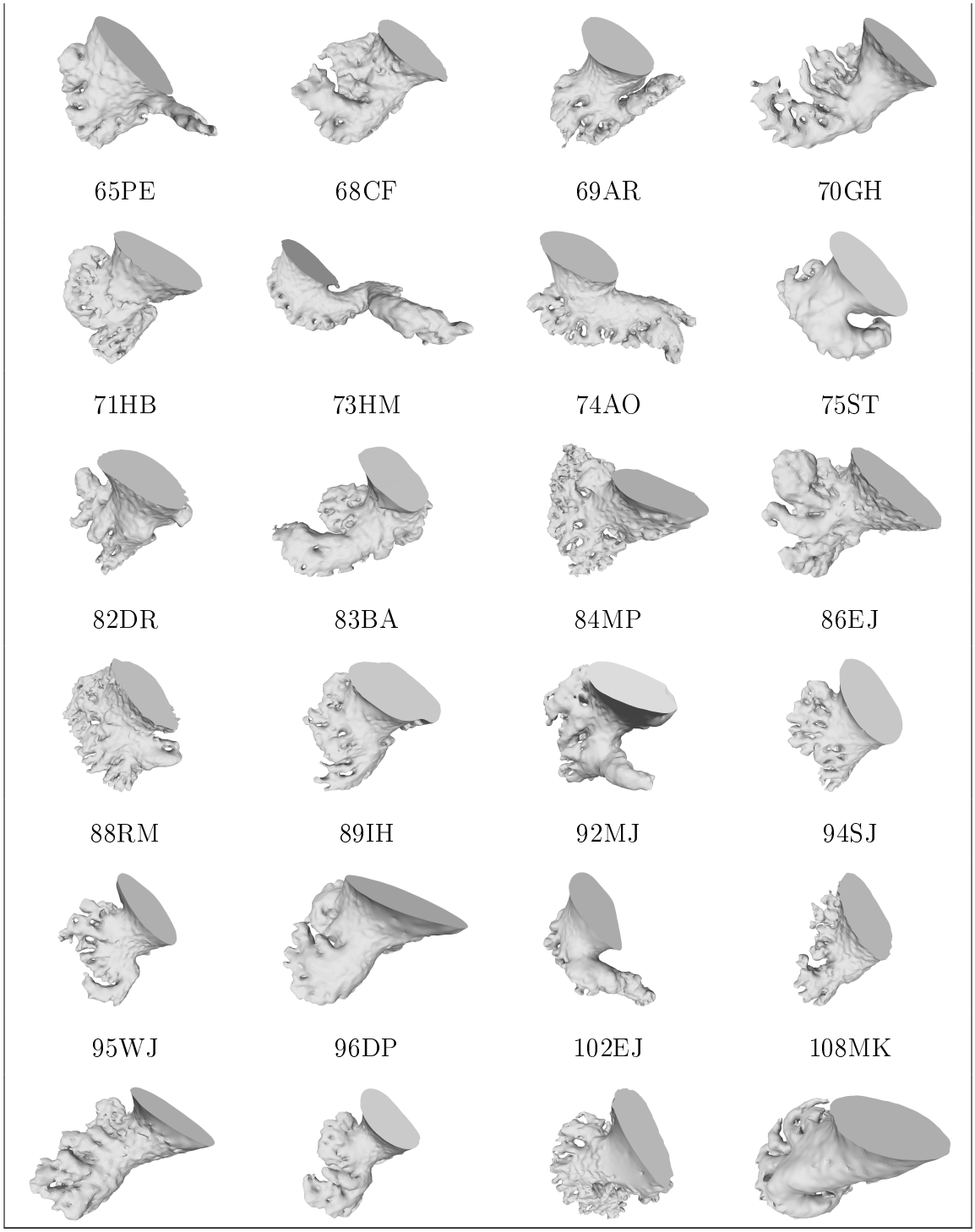
4AI	6HD	7ME	10PA
			
11CA	13MM	17HM	24AM
			
27GA	30SM	31PC	33YE
			
34RS	35AE	36MJ	40ET


42BK
44BM
45DC
46RR
48PA
50MA
51JM
54SD
61MA
62CH
67RA
72VS
77MC
78SE
80AL
81CA
85GP
87HS
90NJ
91PF
93TL
98AA
100DS
101GM


103HT
104BR
105MJ
106WC
110BC
111DS

Cluster B


1BA
2AA
5ML
8BM
9JJ
12LR
15LM
16NC
18SA
19GB
28AG
29AS
38BJ
39MF
41MR
43LM
53SD
56JP
57SL
58BH


65PE
68CF
69AR
70GH
71HB
73HM
74AO
75ST
82DR
83BA
84MP
86EJ
88RM
89IH
92MJ
94SJ
95WJ
96DP
102EJ
108MK

Acknowledgements

When I started my Ph.D. studies in 2020, I would have never imagined that I would be grown so much academically and personally. The merit of it is not only mine. First, they are the merit of the people who accompanied me during my Ph.D. journey.

The first person I must thank is my supervisor Prof. Michele Tumminello. Everything I know about complex systems is his merit.

I enjoyed working at the Department of Economics, Business and Statistics of the University of Palermo. I thank all people who contributed directly and indirectly to my education as a Ph.D. student. First of all, my second supervisor, Prof. Andrea Consiglio, without his encouragement and support, I would have never started this journey. He guided me even before the beginning of my Ph.D, and I have enjoyed working with him and his co-author Stavros A. Zenios.

I also thank the coordinator of the Ph.D. program, Prof. Vito M.R. Muggeo, the Director of the Department, Prof. Angelo Mineo, and all Professors who organised courses and seminars for Ph.D. students. I also thank my colleagues in the Ph.D. program for our lovely time together over these years.

My visiting period at University College of London has been an amazing experience that has changed my life. I thank Dr. Giorgia Bosi, my local supervisor, for his guidance and advice during the six-month visiting period at the Mechanical Engineering Department of the University College of London in the UK. I also thank my Ph.D. colleagues for the interesting projects we are still carrying out together and all the fun we had in London.

I also thank my company supervisors, Dr. Claudia Coronnello and Prof. Gaetano Buriesci, for their advice and disponibility over these years.

My family was a fundamental point of reference during my Ph.D. studies. My father, mother and brother always supported me, even during the toughest times of this journey. I owe them a large debt of gratitude.

Last but not least, I thank my fiancée, Valentina Calderone. Every line of this thesis represents some time we did not spend together. But she is happy anyway, and she is proud of me. I am deeply grateful to her. Every effort of this thesis is dedicated to her.

Bibliography

[1] miRTarBase. https://mirtarbase.cuhk.edu.cn/~miRTarBase/miRTarBase_2022/php/search.php.

[2] N. M. Al-Saady, O. A. Obel, and A. J. Camm. Left atrial appendage: Structure, function, and role in thromboembolism. *Heart*, 82(5):547–554, 1999.

[3] U. Ala. Competing endogenous rnas, non-coding rnas and diseases: an intertwined story. *Cells*, 9(7):1574, 2020.

[4] D. Arthur, S. Vassilvitskii, et al. k-means++: The advantages of careful seeding. In *Soda*, volume 7, pages 1027–1035, 2007.

[5] R.B. Asha and K.R. Suresh Kumar. Credit card fraud detection using artificial neural network. *Global Transitions Proceedings*, 2(1):35–41, 2021.

[6] J.D. Banfield and A.E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821, 1993.

[7] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions*, 53:370–418, 1763.

[8] K. Bedeir, S. Warriner, E. Kofsky, C. Gullett, and B. Ramlawi. Left atrial appendage epicardial clip (atriclip): essentials and post-procedure management. *Journal of atrial fibrillation*, 11(6), 2019.

[9] R. Beigel, N. C. Wunderlich, S. Y. Ho, R. Arsanjani, and R. J. Siegel. The left atrial appendage: Anatomy, function, and noninvasive evaluation. *JACC: Cardiovascular Imaging*, 7(12):1251–1265, 2014.

[10] J.C. Bezdek. Cluster validity with fuzzy sets. 3(3):58–73, 1973.

[11] J.C. Bezdek. Numerical taxonomy with fuzzy sets. *Journal of mathematical biology*, 1(1):57–71, 1974.

[12] J.C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media, 2013.

[13] N. Biggs, E.K. Lloyd, and R.J. Wilson. *Graph Theory, 1736-1936*. Oxford University Press, 1986.

[14] C.M. Bishop. Pattern recognition and machine learning. *Springer google scholar*, 2:645–678, 2006.

[15] R. Blanco, P. Malagón, S. Briongos, and J.M. Moya. Anomaly detection using gaussian mixture probability model to implement intrusion detection system. In *Hybrid Artificial Intelligent Systems: 14th International Conference, HAIS 2019, León, Spain, September 4-6, 2019, Proceedings 14*, pages 648–659, 2019.

[16] V. D. Blondel, J. . Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), 2008.

[17] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, 1992.

[18] G. M. Bosi, A. Cook, R. Rai, L. J. Menezes, S. Schievano, R. Torii, and Burriesci Burriesci, G. Computational fluid dynamic analysis of the left atrial appendage to predict thrombosis risk. *Frontiers in Cardiovascular Medicine*, 5, 2018.

[19] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[20] L. P. Budge, K. M. Shaffer, J. R. Moorman, D. E. Lake, J. D. Ferguson, and J. M. Mangrum. Analysis of in vivo left atrial appendage morphology in patients with atrial fibrillation: A direct comparison of transesophageal echocardiography, planar cardiac ct, and segmented three-dimensional cardiac ct. *Journal of Interventional Cardiac Electrophysiology*, 23(2):87–93, 2008.

[21] E.A. Budisteanu and I.G. Mocanu. Combining supervised and unsupervised learning algorithms for human activity recognition. *Sensors*, 21 (18):6309, 2021.

[22] D.T. Bui, P. Tsangaratos, V.T. Nguyen, N. Van Liem, and P.T. Trinh. Comparing the prediction performance of a deep learning neural network model with conventional machine learning models in landslide susceptibility assessment. *Catena*, 188:104426, 2020.

[23] G.A. Calin and C.M. Croce. Microrna signatures in human cancers. *Nature reviews cancer*, 6(11):857–866, 2006.

[24] R.J.G.B. Campello and E.R. Hruschka. A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems*, 157 (21):2858–2875, 2006.

[25] R. Caruana and A. Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168, 2006.

[26] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern recognition*, 28(5):781–793, 1995.

[27] S. Celi, E. Vignali, K. Capellini, and E. Gasparotti. On the role and effects of uncertainties in cardiovascular in silico analyses. *Frontiers in Medical Technology*, 3, 2021.

[28] C. Chen, D.A. Ridzon, A.J. Broomer, Z. Zhou, D.H. Lee, J.T. Nguyen, M. Barbisin, N.L. Xu, V.R. Mahuvakar, M.R. Andersen, et al. Real-time quantification of micrornas by stem–loop rt–pcr. *Nucleic acids research*, 33(20):e179–e179, 2005.

[29] Ling Chen, Sung-Hao Huang, Tzu-Hsiang Wang, Tzuo-Yun Lan, Vincent Tseng, Hsuan-Ming Tsao, Hsueh-Han Wang, and Gau-Jun Tang. Deep learning-based automatic left atrial appendage filling defects assessment on cardiac computed tomography for clinical and subclinical atrial fibrillation patients. *Heliyon*, 9:e12945, 01 2023.

[30] S.W. Chi, J.B. Zang, A. Mele, and R.B. Darnell. Argonaute hits-clip decodes microrna–mrna interaction maps. *Nature*, 460(7254):479–486, 2009.

[31] A. Clauset, M.E.J. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[32] C. Coronnello and P.V. Benos. Comir: combinatorial microrna target prediction tool. *Nucleic acids research*, 41(W1):W159–W164, 2013.

[33] C. Coronnello, R. Hartmaier, A. Arora, L. Huleihel, K.V. Pandit, A.S. Bais, M. Butterworth, N. Kaminski, G.D. Stormo, S. Oesterreich, et al. Novel modeling of combinatorial mirna targeting identifies snp with potential role in bone density. *PLoS computational biology*, 8(12):e1002830, 2012.

[34] L. Cressa, D. Hildick-Smith, A. Herrara, and F. Ruiz. A new practical anatomical classification for left atrial appendage closure. *EuroIntervention*, 15:88–9, 2019.

[35] M.E. Cusick, N. Klitgord, M. Vidal, and D.E. Hill. Interactome: gateway into systems biology. *Human molecular genetics*, 14(suppl_2):R171–R181, 2005.

[36] G. Cybenko, D.P. O'Leary, and J. Rissanen. *The mathematics of information coding, extraction and distribution*, volume 107. Springer Science & Business Media, 1998.

[37] R.N. Dave. Validating fuzzy partitions obtained through c-shells clustering. *Pattern recognition letters*, 17(6):613–623, 1996.

[38] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

[39] Y. Delpu, D. Larrieu, M. Gayral, D. Arvanitis, M. Dufresne, P. Cordelier, and J. Torrisani. Noncoding rnas: clinical and therapeutic applications. In *Drug discovery in Cancer epigenetics*, pages 305–326. 2016.

[40] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

[41] L. Di Biase, P. Santangeli, M. Anselmino, P. Mohanty, I. Salvetti, S. Gili, R. Horton, J. E. Sanchez, R. Bai, S. Mohanty, A. Pump,

M. Cereceda Brantes, G. J. Gallinghouse, J. D. Burkhardt, F. Cesarani, M. Scaglione, A. Natale, and F. Gaita. Does the left atrial appendage morphology correlate with the risk of stroke in patients with atrial fibrillation? results from a multicenter study. *Journal of the American College of Cardiology*, 60(6):531–538, 2012.

[42] R. Donnino, P.A. Tunick, and I. Kronzon. Left atrial appendage thrombus outside of a 'successful'ligation. *European Journal of Echocardiography*, 9(3):397–398, 2008.

[43] J.C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.

[44] A. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. Marks. Microrna targets in drosophila. *Genome biology*, 4:1–27, 2003.

[45] G. Ernst, C. Stöllberger, F. Abzieher, W. Veit-Dirscherl, E. Bonner, B. Bibus, B. Schneider, and J. Slany. Morphology of the left atrial appendage. *The Anatomical Record*, 242(4):553–561, 1995.

[46] A.E. Erson and E.M. Petty. Micrornas in development and disease. *Clinical genetics*, 74(4):296–306, 2008.

[47] A.E. Erson-Bensan. Introduction to micrornas in biological systems. *MiRNomics: MicroRNA biology and computational analysis*, pages 1–14, 2014.

[48] A. Fitzgibbon, M. Pilu, and R. B. Fisher. Direct least square fitting of ellipses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):476–480, 1999.

[49] E. Fix and J.L. Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review / Revue Internationale de Statistique*, 57(3):238–247, 1989.

[50] C. Fraley and A.E. Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The computer journal*, 41(8):578–588, 1998.

[51] L.C. Freeman et al. Centrality in social networks: Conceptual clarification. *Social network: critical concepts in sociology. Londres: Routledge*, 1:238–263, 2002.

[52] R.C. Friedman, K.K.H. Farh, C.B. Burge, and D.P. Bartel. Most mammalian mrnas are conserved targets of microrna. *Genome research*, 19 (1):92–105, 2009.

[53] B.F. Gage, C. Van Walraven, L. Pearce, R.G. Hart, P.J. Koudstaal, B.S.P. Boode, and P. Petersen. Selecting patients with atrial fibrillation for anticoagulation: stroke risk stratification in patients taking aspirin. *Circulation*, 110(16):2287–2292, 2004.

[54] M. Gagliardi and M.R. Matarazzo. Rip: Rna immunoprecipitation. *Polycomb Group Proteins: Methods and Protocols*, pages 73–86, 2016.

[55] A.M Gillinov, G. Pettersson, and D.M. Cosgrove. Stapled excision of the left atrial appendage. *The Journal of Thoracic and Cardiovascular Surgery*, 129(3):679–680, 2005.

[56] M. Girvan and M.E.J. Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99 (12):7821–7826, 2002.

[57] M. E. Goldman, L. A. Pearce, R. G. Hart, M. Zabalgoitia, R. W. Asinger, R. Safford, and J. L. Halperin. Pathophysiologic correlates of thromboembolism in nonvalvular atrial fibrillation: I. reduced flow velocity in the left atrial appendage (the stroke prevention in atrial fibrillation [spaf-iii] study). *Journal of the American Society of Echocardiography*, 12(12):1080–1087, 1999.

[58] C. Goutte and E. Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *Advances in Information Retrieval*, pages 345–359. Springer Berlin Heidelberg, 2005.

[59] J. C. Gower and G. J. S. Ross. Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 18(1):54–64, 1969.

[60] T.Y. Ha. Micrornas in human diseases: from cancer to cardiovascular disease. *Immune network*, 11(3):135, 2011.

[61] M. Hafner, P. Landgraf, J. Ludwig, A. Rice, T. Ojo, C. Lin, D. Holoch, C. Lim, and T. Tuschl. Identification of micrornas and other small regulatory rnas using cdna library sequencing. *Methods*, 44(1):3–12, 2008.

[62] J.A. Hanley and B.J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[63] R.G. Hart, L.A. Pearce, and M.I. Aguilar. Meta-analysis: Antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation. *Annals of Internal Medicine*, 146(12):857–867, 2007.

[64] J.A. Hartigan and M.A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[65] John A Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

[66] E. K. Heist, M. Refaat, S. B. Danik, G. Holmvang, J. N. Ruskin, and M. Mansour. Analysis of the left atrial appendage by magnetic resonance angiography in patients with atrial fibrillation. *Heart Rhythm*, 3(11): 1313–1318, 2006.

[67] D.G. Hendrickson, D.J. Hogan, D. Herschlag, J.E. Ferrell, and P.O. Brown. Systematic identification of mrnas recruited to argonaute 2 by specific micrornas and corresponding changes in transcript abundance. *PloS one*, 3(5):e2126, 2008.

[68] T.K. Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282, 1995.

[69] G.T. Huang, C. Athanassiou, and P.V. Benos. mirconnx: condition-specific mrna-microrna network integrator. *Nucleic acids research*, 39 (suppl_2):W416–W423, 2011.

[70] J.C. Huang, Q.D. Morris, and B.J. Frey. Bayesian inference of microrna targets from sequence and expression data. *Journal of Computational Biology*, 14(5):550–563, 2007.

[71] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[72] C. Jin, J. Feng, L. Wang, H. Yu, J. Liu, J. Lu, and J. Zhou. Left atrial appendage segmentation using fully convolutional neural networks and modified three-dimensional conditional random fields. *IEEE Journal of Biomedical and Health Informatics*, 22(6):1906–1916, 2018.

[73] W.D. Johnson, A.K. Ganjoo, C.D. Stone, R.C. Srivyas, and M. Howard. The left atrial appendage: our most lethal human attachment! surgical implications. *European journal of cardio-thoracic surgery*, 17(6):718–722, 2000.

[74] R. Kaminski, A. Kosinski, M. Brala, G. Piwko, E. Lewicka, A. Dabrowska-Kugacka, G. Raczak, and D. Kozlowski. Grzybiak met al. variability of the left atrial appendage in human hearts. *PLoS One*, 10:e0141901, 2015.

[75] G. Karlebach and R. Shamir. Modelling and analysis of gene regulatory networks. *Nature reviews Molecular cell biology*, 9(10):770–780, 2008.

[76] L. Kaufman and P.J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[77] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal. The role of site accessibility in microrna target recognition. *Nature genetics*, 39(10): 1278–1284, 2007.

[78] T. Kimura, S. Takatsuki, K. Inagawa, Y. Katsumata, T. Nishiyama, N. Nishiyama, K. Fukumoto, Y. Aizawa, Y. Tanimoto, K. Tanimoto, M. Jinzaki, and K. Fukuda. Anatomical characteristics of the left atrial appendage in cardiogenic stroke with low chads2 scores. *Heart Rhythm*, 10(6):921–925, 2013.

[79] T. Kimura, S. Takatsuki, S. Miyoshi, Y. Katsumata, T. Nishiyama, N. Nishiyama, Y. Tanimoto, Y. Aizawa, M. Jinzaki, and K. Fukuda. Pericardial endoscopy-guided left atrial appendage ligation: A pilot study in a canine model. *Circulation: Cardiovascular Interventions*, 7(6):844–850, 2014.

[80] Y. Kohavi and H. Davdovich. Topological dimensions, hausdorff dimensions & fractals. *Bar-llan University*, pages 1–16, 2006.

[81] I. Kononenko. Comparison of inductive and naive bayesian learning approaches to automatic knowledge acquisition. *Current trends in knowledge acquisition*, 8:190, 1990.

[82] M. Koplay, C. Erol, Y. Paksoy, A.S. Kivrak, and S. Özbek. An investigation of the anatomical variations of left atrial appendage by multidetector computed tomographic coronary angiography. *European journal of radiology*, 81(7):1575–1580, 2012.

[83] C. Kwak and A. Clayton-Matthews. Multinomial logistic regression. *Nursing research*, 51(6):404–410, 2002.

[84] J.M. Lacomis, O. Goitein, C. Deible, P.L. Moran, G. Mamone, S. Madan, and D. Schwartzman. Dynamic multidimensional imaging of the human left atrial appendage. *Europace*, 9(12):1134–1140, 2007.

[85] C.J. Lee and J. Ansell. Direct thrombin inhibitors. *British Journal of Clinical Pharmacology*, 72:581–592, 2011.

[86] C. Lei, Q. Gao, R. Wei, Q. Li, X. Liu, L. Wu, Y. Yao, H. Fan, and Z. Zheng. Fractal geometry illustrated left atrial appendage morphology that predicted thrombosis and stroke in patients with atrial fibrillation. *Frontiers in Cardiovascular Medicine*, 9, 2022.

[87] E.A. Leicht and M.E.J. Newman. Community structure in directed networks. *Physical review letters*, 100(11):118703, 2008.

[88] C. Li, B. Gao, X. Liu, Q. Fan, X. Zhang, G. Liu, H. Yang, P. Feng, Y. Wang, and P. Song. Quantitative evaluation of the substantially variable morphology and function of the left atrial appendage and its relation with adjacent structures. *PLoS ONE*, 10(7), 2015.

[89] G. Lippi, F. Sanchis-Gomar, and G. Cervellin. Global epidemiology of atrial fibrillation: an increasing epidemic and public health challenge. *International journal of stroke*, 16(2):217–221, 2021.

[90] C.G. Liu, G.A. Calin, S. Volinia, and C.M. Croce. Microrna expression profiling using microarrays. *Nature protocols*, 3(4):563–578, 2008.

[91] D.J. Livingstone, D.T. Manallack, and I.V. Tetko. Data modelling with neural networks: advantages and limitations. *Journal of computer-aided molecular design*, 11:135–142, 1997.

[92] S. Lloyd. Least squares quantization in pcm. *IEEE Trans. Info. Theor.*, 28(2):129–137, 1982.

[93] J.R. López-Mínguez, R. González-Fernández, C. Fernández-Vegas, V. Millán-Nuñez, M.E. Fuentes-Cañamero, J.M. Nogales-Asensio, J. Doncel-Vecino, J. Elduayen-Gragera, S.Y. Ho, and D. Sánchez-Quintana. Anatomical classification of left atrial appendages in specimens applicable to ct imaging techniques for implantation of amplatzer cardiac plug. *Journal of cardiovascular electrophysiology*, 25(9):976–984, 2014.

[94] J. MacQueen. Some methods for classification and analysis of multivariate observations. volume 1, pages 281–297, 1967.

[95] B.B. Mandelbrot and B.B. Mandelbrot. *The fractal geometry of nature*, volume 1. WH freeman New York, 1982.

[96] T. A. Manolio, R. A. Kronmal, G. L. Burke, D. H. O'Leary, and T. R. Price. Short-term predictors of incident stroke in older adults: The cardiovascular health study. *Stroke*, 27(9):1479–1486, 1996.

[97] A. Masci, L. Barone, L. Dedè, M. Fedele, C. Tomasi, A. Quarteroni, and C. Corsi. The impact of left atrium appendage morphology on stroke risk assessment in atrial fibrillation: A computational fluid dynamics study. *Frontiers in Physiology*, 9, 2019.

[98] B.W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

[99] F. Maturo and R. Verde. Combining unsupervised and supervised learning techniques for enhancing the performance of functional data classifiers. *Computational Statistics*, pages 1–38, 2022.

[100] W. Mcculloch and W. Pitts. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:127–147, 1943.

[101] S.E. McGeary, K.S. Lin, C.Y. Shi, T.M. Pham, N. Bisaria, G.M. Kelley, and D.P. Bartel. The biochemical basis of microrna targeting efficacy. *Science*, 366(6472):eaav1741, 2019.

[102] S. Möbius-Winkler, M. Sandri, N. Mangner, P. Lurz, I. Dähnert, and G. Schuler. The watchman left atrial appendage closure device for atrial fibrillation. *JoVE (Journal of Visualized Experiments)*, (60):e3671, 2012.

[103] K.L. Moore and T.V.N. Persaud. *The Developing Human: Clinically Oriented Embryology*. Elsevier,Saunders, 10 edition, 2016.

[104] A. Muniategui, R. Nogales-Cadenas, M. Vázquez, X. L. Aranguren, X. Agirre, A. Luttun, F. Prosper, A. Pascual-Montano, and A. Rubio. Quantification of mirna-mrna interactions. *PloS one*, 7(2):e30766, 2012.

[105] G. Musotto, A. Monteleone, D. Vella, S. Di Leonardo, A. Viola, G. Pitarresi, B. Zuccarello, A. Pantano, A. Cook, G.M. Bosi, et al. The role of patient-specific morphological features of the left atrial appendage on the thromboembolic risk under atrial fibrillation. *Frontiers in Cardiovascular Medicine*, 9:894187, 2022.

[106] G. Musotto, A. Monteleone, D. Vella, B. Zuccarello, R. Cannova, A. Cook, G.M. Bosi, and G. Burriesci. Fluid-structure interaction analysis of the thromboembolic risk in the left atrial appendage under atrial fibrillation: Effect of hemodynamics and morphological features. *Computer Methods and Programs in Biomedicine*, 246:108056, 2024.

[107] S. Nam, M. Li, K. Choi, C. Balch, S. Kim, and K.P. Nephew. Microrna and mrna integrated analysis (mmia): a web tool for examining biological functions of microrna expression. *Nucleic acids research*, 37(suppl_2): W356–W362, 2009.

[108] A. Newell, H.A. Simon, et al. *Human problem solving*, volume 104. Prentice-hall Englewood Cliffs, NJ, 1972.

[109] M.E.J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006.

[110] M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[111] J. A. Odell, J. L. Blackshear, E. Davies, W. J. Byrne, C. F. Kollmorgen, W. D. Edwards, and T. A. Orszulak. Thoracoscopic obliteration of the left atrial appendage: Potential for stroke reduction? *Annals of Thoracic Surgery*, 61(2):565–569, 1996.

[112] T. Ohtsuka, M. Ninomiya, T. Nonaka, M. Hisagi, T. Ota, and T. Mizutani. Thoracoscopic stand-alone left atrial appendectomy for thromboembolism prevention in nonvalvular atrial fibrillation. *Journal of the American College of Cardiology*, 62(2):103–107, 2013.

[113] N. Otsu. Threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern*, SMC-9(1):62–66, 1979.

[114] A.F. Palazzo and T.R. Gregory. The case for junk dna. *PLOS Genetics*, 10(5):1–8, 2014.

[115] J.W. Park, A. Bethencourt, H. Sievert, G. Santoro, B. Meier, K. Walsh, J.R. Lopez-Minquez, D. Meerkin, M. Valdés, O. Ormerod, et al. Left atrial appendage closure with amplatzer cardiac plug in atrial fibrillation: initial european experience. *Catheterization and Cardiovascular Interventions*, 77(5):700–706, 2011.

[116] S.M. Prigarin, K. Sandau, M. Kazmierczak, and K. Hahn. Estimation of fractal dimension: A survey with numerical experiments and software description. *International Journal of Biomathematics and Biostatistics*, 2(1):167–180, 2013.

[117] R.C. Prim. Shortest connection networks and some generalizations. *The Bell System Technical Journal*, 36(6):1389–1401, 1957.

[118] M. Prince, R. Bryce, E. Albanese, A. Wimo, W. Ribeiro, and C. P. Ferri. The global prevalence of dementia: A systematic review and metaanalysis. *Alzheimer's and Dementia*, 9(1):63–75.e2, 2013.

[119] P. Rajiah, M. Alkhouli, J. Thaden, T. Foley, E. Williamson, and P. Ranganath. Pre- and postprocedural ct of transcatheter left atrial appendage closure devices. *Radiographics*, 41(3):680–698, 2021.

[120] D.A. Reynolds. *A Gaussian mixture modeling approach to text-independent speaker identification*. Georgia Institute of Technology, 1992.

[121] M. Rosvall and C.T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences*, 105(4):1118–1123, 2008.

[122] M. Rosvall, D. Axelsson, and C.T. Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23, 2009.

[123] P.J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[124] T.W. Sadler. *Langman's Medical Embryology*. Wolters Kluwer Health, 2022.

[125] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A bayesian approach to filtering junk e-mail. In *Learning for Text Categorization: Papers from the 1998 workshop*, volume 62, pages 98–105, 1998.

[126] G. Sales, A. Coppe, A. Bisognin, M. Biasiolo, S. Bortoluzzi, and C. Romualdi. Magia, a web-based tool for mirna and genes integrated analysis. *Nucleic acids research*, 38(suppl_2):W352–W359, 2010.

[127] A.L. Samuel. *Some Studies in Machine Learning Using the Game of Checkers. I*, pages 335–365. Springer New York, 1988.

[128] B. Schwikowski, P. Uetz, and S. Fields. A network of protein-protein interactions in yeast. *Nature biotechnology*, 18(12):1257–1261, 2000.

[129] L. Scrucca, M. Fop, T.B. Murphy, and A.E. Raftery. mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *The R journal*, 8(1):289, 2016.

[130] S.A. Shabalina and E.V. Koonin. Origins and evolution of eukaryotic rna interference. *Trends in ecology & evolution*, 23(10):578–587, 2008.

[131] L. Shapira, A. Shamir, and D. Cohen-Or. Consistent mesh partitioning and skeletonisation using the shape diameter function. *Visual Computer*, 24(4):249–259, 2008.

[132] J. Shirani and J. Alaeddini. Structural remodeling of the left atrial appendage in patients with chronic non-valvular atrial fibrillation: Implications for thrombus formation, systemic embolism, and assessment

by transesophageal echocardiography. *Cardiovascular Pathology*, 9(2): 95–101, 2000.

[133] T.G. Smith Jr, G.D. Lange, and W.B. Marks. Fractal methods and results in cellular morphology—dimensions, lacunarity and multifractals. *Journal of neuroscience methods*, 69(2):123–136, 1996.

[134] R.R. Sokal and C.D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas science bulletin*, 38:1409–1438, 1958.

[135] J. Steffel, R. Collins, M. Antz, P. Cornu, L. Desteghe, K.G. Haeusler, J. Oldgren, H. Reinecke, V. Roldan-Schilling, N. Rowell, et al. 2021 european heart rhythm association practical guide on the use of non-vitamin k antagonist oral anticoagulants in patients with atrial fibrillation. *EP Europace*, 23(10):1612–1676, 2021.

[136] S.V. Stehman. Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 62(1):77–89, 1997.

[137] P. Su, K.P. Mccarthy, and S.Y. Ho. Occluding the left atrial appendage: anatomical considerations. *Heart*, 94(9):1166–1170, 2008.

[138] K. Słodowska, E. Szczepanek, D. Dudkiewicz, J. Hołda, F. Bolechała, M. Strona, M. Lis, J. Batko, M. Koziej, and M. K. Hołda. Morphology of the left atrial appendage: Introduction of a new simplified shape-based classification system. *Heart Lung and Circulation*, 30(7):1014–1022, 2021.

[139] A. Tagliasacchi, I. Alhashim, M. Olson, and H. Zhang. Mean curvature skeletons. *Computer Graphics Forum*, 31(5):1735–1744, 2012.

[140] Itaru Takayashiki, Akio Doi, Toru Kato, Hiroki Takahashi, Shoto Sekimura, Maiko Hozawa, and Yoshihiro Morino. Method for left atrial appendage segmentation using heart ct images. In *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*, pages 1–5, 2019.

[141] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[142] V.A. Traag, L. Waltman, and N.J. Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233, 2019.

[143] S. Tualu. Fractal analysis of normal retinal vascular network. *Oftalmologia (Bucharest, Romania: 1990)*, 55(4):11–16, 2011.

[144] B. Tuccillo, O. Stümper, J. Hess, R. J. Van Suijlen, E. Bos, J. R. T. C. Roelandt, and G. R. Sutherland. Transoesophageal echocardiographic evaluation of atrial morphology in children with congenital heart disease. *European heart journal*, 13(2):223–231, 1992.

[145] M. Tumminello, T. Aste, T. Di Matteo, and R.N. Mantegna. A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences*, 102(30):10421–10426, 2005.

[146] M. Tumminello, C. Coronnello, F. Lillo, S. Micciche, and R.N. Mantegna. Spanning trees and bootstrap reliability estimation in correlation-based networks. *International Journal of Bifurcation and Chaos*, 17(07):2319–2329, 2007.

[147] C.J. van Rijsbergen. *Information retrieval*. Butterworth-Heinemann, 1979.

[148] J. P. Veinot, P. J. Harrity, F. Gentile, B. K. Khandheria, K. R. Bailey, J. T. Eickholt, J. B. Seward, A. J. Tajik, and W. D. Edwards. Anatomy of the normal left atrial appendage: A quantitative study of age-related changes in 500 autopsy hearts: Implications for echocardiographic examination. *Circulation*, 96(9):3112–3115, 1997.

[149] L.C. Villaruz, G. Huang, M. Romkes, J.M. Kirkwood, S.C. Buch, T. Nukui, K.T. Flaherty, S.J. Lee, M.A. Wilson, K.L. Nathanson, et al. Microrna expression profiling predicts clinical outcome of carboplatin/paclitaxel-based therapy in metastatic melanoma treated on the ecog-acrin trial e2603. *Clinical epigenetics*, 7:1–9, 2015.

[150] F. Wang, M. Zhu, X. Wang, W. Zhang, Y. Su, Y. Lu, X. Pan, D Gao, X. Zhang, W. Chen, et al. Predictive value of left atrial appendage lobes on left atrial thrombus or spontaneous echo contrast in patients with non-valvular atrial fibrillation. *BMC cardiovascular disorders*, 18:1–11, 2018.

[151] X. Wang and I.M. El Naqa. Prediction of both conserved and nonconserved microrna targets in animals. *Bioinformatics*, 24(3):325–332, 2008.

[152] Y. Wang, L. Di Biase, R. P. Horton, T. Nguyen, P. Morhanty, and A. Natale. Left atrial appendage studied by computed tomography to help planning for appendage closure device placement. *Journal of cardiovascular electrophysiology*, 21(9):973–982, 2010.

[153] J.H.Jr Ward. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[154] Stanley Wasserman and Katherine Faust. Cambridge university press, 1994.

[155] A. Widodo and B. Yang. Application of nonlinear feature extraction and support vector machines for fault diagnosis of induction motors. *Expert Systems with Applications*, 33(1):241–250, 2007.

[156] B. Wightman, I. Ha, and G. Ruvkun. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in c. elegans. *Cell*, 75(5):855–862, 1993.

[157] W. Winterbach, P.V. Mieghem, M. Reinders, H. Wang, and D.de Ridder. Topology of molecular interaction networks. *BMC systems biology*, 7:1–15, 2013.

[158] L. Wu, E. Liang, S. Fan, L. Zheng, Z. Du, S. Liu, F. Hu, X. Fan, G. Chen, L. Ding, and Y. Yao. Relation of left atrial appendage morphology determined by computed tomography to prior stroke or to increased risk of stroke in patients with atrial fibrillation. *American Journal of Cardiology*, 123(8):1283–1286, 2019.

[159] S. Wuchty and E. Almaas. Evolutionary cores of domain co-occurrence networks. *BMC Evolutionary Biology*, 5, 2005.

[160] S. Yaghi, A. D. Chang, R. Akiki, S. Collins, T. Novack, M. Hemendinger, A. Schomer, B. M. Grory, S. Cutting, T. Burton, C. Song, A. Poppas, R. McTaggart, M. Jayaraman, A. Merkler, H. Kamel, M. S. V. Elkind, K. Furie, and M. K. Atalay. The left atrial appendage morphology is associated with embolic stroke subtypes using a simple classification sys-

tem: A proof of concept study. *Journal of Cardiovascular Computed Tomography*, 14(1):27–33, 2020.

[161] M. Yamamoto, Y. Seo, N. Kawamatsu, K. Sato, A. Sugano, T. Machino-Ohtsuka, R. Kawamura, H. Nakajima, M. Igarashi, Y. Sekiguchi, T. Ishizu, and K. Aonuma. Complex left atrial appendage morphology and left atrial appendage thrombus formation in patients with atrial fibrillation. *Circulation: Cardiovascular Imaging*, 7(2):337–343, 2014.

[162] M. Yousef, S. Jung, A.V. Kossenkov, L.C. Showe, and M.K. Showe. Naïve bayes for microrna target predictions-machine learning for microrna targets. *Bioinformatics*, 23(22):2987–2992, 2007.

La borsa di dottorato è stata cofinanziata con risorse del
Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 (CCI 2014IT16M2OP005),
Fondo Sociale Europeo, Azione I.1 “Dottorati Innovativi con caratterizzazione Industriale”



UNIONE EUROPEA
Fondo Sociale Europeo



