UNIVERSITA' POLITECNICA DELLE MARCHE

FACOLTA' DI INGEGNERIA

Master's Degree in Biomedical Engineering

Curriculum: E-Health and Clinical Engineering

Machine Learning and Analysis of Clinical and Electrophysiological Data in Patients with Ventricular Arrhythmias: Potential Prognostic Role

Supervisor:

Prof. **Francesco Piva**

Candidate:

Paolo Veri

Co-supervisor:

Prof. **Michela Casella**

Academic Year: 2022 / 2023

INDEX

1. INTRODUCTION **05**

1.1. ARRHYTHMIAS 05

1.1.1. ANATOMY OF THE CARDIAC CONDUCTION SYSTEM 05

1.1.2. CARDIAC ELECTROPHYSIOLOGY 07

1.1.3. NORMAL CARDIAC RHYTHM 09

1.1.4. MECHANISM OF TYPICAL REENTRY 10

1.2. CAUSES OF HEART FAILURE 12

1.3. CARDIOMYOPATHIES 14

1.3.1. DILATED CARDIOMYOPATHY 15

1.3.2. HYPERTROPHIC CARDIOMYOPATHY 15

1.3.3. RESTRICTIVE CARDIOMYOPATHY 16

1.4. INCIDENCE AND MORTALITY 16

1.5. TREATMENT 18

1.5.1. IMPLANTABLE CARDIAC DEFIBRILLATOR (ICD) 18

1.5.2. ABLATION 18

1.5.2.1. IMPORTANCE OF MAPPING 19

1.5.2.2. PROCEDURE 20

1.6. POTENTIAL PREDICTORS 20

2. OBJECTIVES **22**

3. MATERIALS AND METHODS **23**

3.1. HARDWARE AND SOFTWARE DEVICES 23

3.2. ELECTROANATOMICAL MAPPING SYSTEM 23

3.2.1. BEGINNING 24

3.2.2. OPERATION 25

3.2.3. EXPORT 28

3.2.4. EXTRACTION ALGORITHM 29
3.2.5. APPLICATION DIFFUSION 32
3.2.5.1. FOR MATLAB USERS 32
3.2.5.2. DESKTOP APPLICATION 33
3.2.5.3. ONLINE VERSION 33
3.2.5.4. LIBRARY FOR OTHER SOFTWARE 34
3.2.5.5. PROJECT SHARING 35
3.3. STUDY COHORT 36
3.4. CLINICAL TESTS 36
3.5. DATA PREPARATION 36
3.6. TRAINING, VALIDATION AND TESTING 40
3.7. MACHINE LEARNING METHODS 41
3.7.1. REGRESSION 41
3.7.2. SUPPORT VECTOR MACHINE (SVM) 43
3.7.3. ARTIFICIAL NEURAL NETWORK (ANN) 49
4. RESULTS 52
4.1. LINEAR 53
4.1.1. LOGISTIC REGRESSION 53
4.1.2. LINEAR SVM 55
4.2. NON-LINEAR 57
4.2.1. KERNEL SVM 57
4.2.2. ARTIFICIAL NEURAL NETWORK 59
5. DISCUSSION 61
6. CONCLUSIONS 66
7. BIBLIOGRAPHY 68

1. INTRODUCTION

1.1. ARRHYTHMIAS

Arrhythmias are defined as changes in the heart rhythm. A normal heart beats regularly and in a coordinated manner because electrical impulses, generated and propagated by muscle cells with specific electrical properties, trigger a series of organized heart muscle contractions. Arrhythmias and conduction disorders are caused by abnormalities in the formation and/or conduction of these electrical impulses. Any heart disease, including structural or functional congenital heart disease, can be associated with arrhythmias. Systemic factors that may cause or contribute to arrhythmias include: electrolyte imbalances, hypoxia, hormonal imbalances, medications, and toxins. All the following information regarding cardiac arrhythmias has been inspired and taken from the MSD manual [1].

1.1.1. ANATOMY OF THE CARDIAC CONDUCTION SYSTEM

There is a group of cells at the junction of the superior vena cava and the upper lateral part of the right atrium, called the sinoatrial node or sinoatrial node. The sinoatrial node generates the first electrical impulse of each normal heartbeat. The discharges from these pacemaker cells propagate to neighboring cells, thereby sequentially stimulating successive regions of the heart. Impulses are transmitted through the atria to the atrioventricular node via preferentially conducting internodal pathways and nonspecialized atrial myocytes. The atrioventricular node is located on the right side of the interatrial septum. It has a slower conduction velocity, which delays the transmission of impulses from the atria to the ventricles. The conduction time of impulses through the atrioventricular node depends on the heart rate, and it is regulated by autonomic tone and circulating catecholamines to maximize cardiac output at any given atrial rate. Except for the anteroseptal region, in which the atria are electrically isolated from the ventricles by the annulus fibrosus. Here, the bundle of His is a continuation of the atrioventricular node and passes through the upper part of the interventricular septum, where it bifurcates into left and right branches, which in turn terminate in Purkinje fibers. The right branch carries impulses to the apex and anterior endocardial region of the right ventricle. The left branch radiates in a fan-shape on the left side of the interventricular septum. Its anterior part (left anterior fasciculus) and posterior part (left posterior fasciculus) stimulate the left side of the interventricular septum, which is

the first part of the ventricle to be electrically activated. The interventricular septum then depolarizes from left to right, followed by an almost simultaneous activation of both the ventricles from the endocardium, through the ventricular wall, to the epicardium. To trace the path of an electrical pulse, see Figure 1.

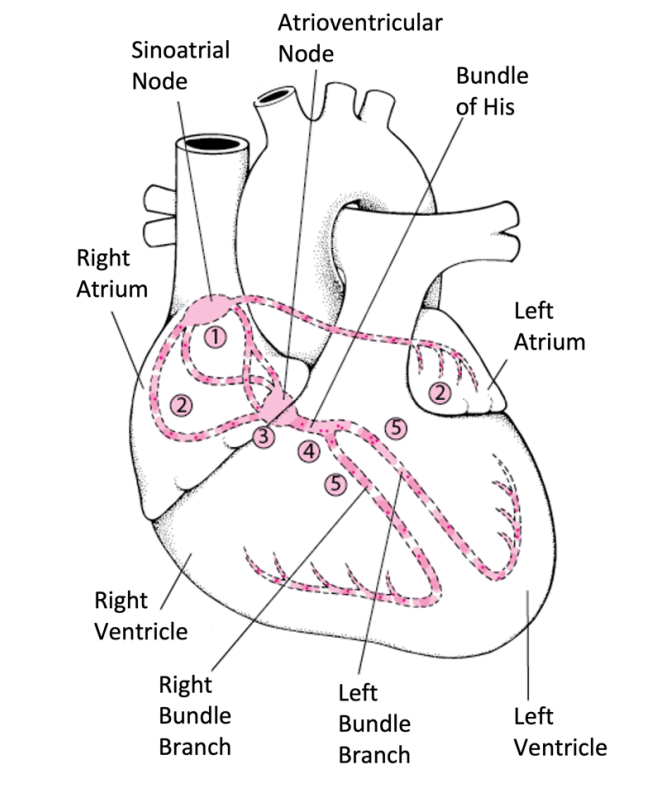

Figure 1. | *An anatomical image of the cardiac conduction system, highlighting key points where the signal originates and is conducted.*

The electrical signal begins in the sinoatrial (SA) node, prompting the contraction of both the right and left atria. As it reaches the atrioventricular (AV) node, there is a brief delay. Afterwards, the signal travels through the bundle of His, which then splits into the right bundle branch, extending to the right ventricle, and into the left bundle branch, leading to the left ventricle. Subsequently, the impulse spreads through the ventricles, resulting in their contraction.

1.1.2. CARDIAC ELECTROPHYSIOLOGY

To comprehend cardiac rhythm disorders, it's essential to have a good grasp of normal cardiac physiology. The passage of ions through the cell membrane of myocytes is meticulously regulated by specific ion channels. These channels orchestrate a cyclic process of depolarization and repolarization within the cell, known as an action potential. Figure 2.

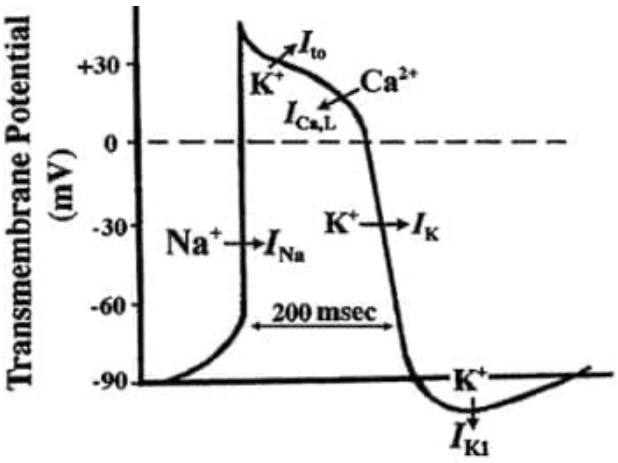

Figure 2. *| Depiction of the action potential, its values in millivolts (mV), and ion movements.*

The action potential in a healthy myocyte get started when the cell's transmembrane potential shifts from its diastolic level of around -90 mV to approximately -50 mV. At this threshold, voltage-gated sodium channels open, leading to a swift depolarization driven by the inflow of sodium ions along their steep concentration gradient. Soon after, these sodium channels inactivate, ceasing the sodium influx. Meanwhile, other time- and voltage-gated ion channels open, enabling calcium to enter through slow calcium channels (contributing to depolarization), and potassium to exit through potassium channels (contributing to repolarization). These processes initially balance each other, sustaining a positive transmembrane potential and prolonging the plateau phase of the action potential. During this phase, calcium's entry into the cell is responsible for electromechanical coupling and myocyte contraction. As calcium entry subsides, potassium outflow increases, rapidly repolarizing the cell until it reaches its resting transmembrane potential of -90 mV. When the cell is depolarized, it becomes refractory to subsequent depolarizations. Initially, it cannot undergo another depolarization (absolute refractory period), and after partial repolarization, a subsequent depolarization is possible but slow (relative refractory period). There are two types of heart tissue:

• Tissues with fast channels
• Tissues with slow channels

Tissues with fast channels, such as atrial and ventricular conduction tissue and the His-Purkinje system, possess a high density of fast sodium channels. Their action potentials are characterized by:

• Little to no spontaneous diastolic depolarization, resulting in very slow rates of pacemaker activity
• Very rapid initial depolarization rates, leading to rapid conduction velocity
• Loss of refractoriness coinciding with repolarization, resulting in short refractory periods and the ability to conduct repetitive impulses at high frequencies

Tissues with slow channels, like the sinoatrial and atrioventricular nodes, have a low density of fast sodium channels, and their action potentials are characterized by:

• Faster spontaneous diastolic depolarization, resulting in a faster rate of pacemaker activity
• Slow initial depolarization rates, leading to slow conduction velocity
• Delayed loss of refractoriness after repolarization, resulting in long refractory periods and the inability to conduct repetitive impulses at high frequencies

Under normal circumstances, the sinoatrial node exhibits the highest frequency of spontaneous diastolic depolarization, making its cells responsible for producing more frequent spontaneous action potentials than other tissues. Thus, in a healthy heart, the sinoatrial node serves as the dominant automatic tissue (pacemaker). If the sinoatrial node fails to generate impulses, the atrioventricular node, typically having the second-highest rate of spontaneous diastolic depolarization, takes over as the pacemaker. Sympathetic stimulation increases the firing rate of pacemaker tissue, while parasympathetic stimulation decreases it. An inwardly directed calcium/potassium current, known as the "funny current," flows through hyperpolarization-activated cyclic nucleotide channels (HCN channels) in sinus node cells, contributing significantly to their automaticity. Inhibiting this current prolongs the time required for critical spontaneous depolarization in pacemaker cells, leading to a reduced heart rate.

1.1.3. NORMAL CARDIAC RHYTHM

The resting sinus heart rate in adults typically falls within the range of 60 to 100 beats per minute. Lower rates, known as sinus bradycardia, are common in young individuals, especially athletes, and also during sleep. On the contrary, higher rates, or sinus tachycardia, can occur during physical exercise, illness, or intense stress, driven by sympathetic activation and the influence of circulating catecholamines. Normally, the heart rate exhibits significant variability throughout the day, often with lower rates observed in the morning prior to waking. A slight increase in heart rate during inspiration, followed by a decrease during expiration (referred to as respiratory sinus arrhythmia), is considered normal. This phenomenon is attributed to fluctuations in vagal tone and is particularly prevalent in healthy young individuals. While it tends to decrease with age, respiratory sinus arrhythmia never entirely disappears. A consistently steady sinus heart rate is abnormal and it is typically observed in patients with autonomic denervation, or in cases of severe cardiac disorders that reduce parasympathetic tone and activate sympathetic tone. As a result, heart rate variability measurements have been suggested as useful indicators of cardiovascular health. Most of the cardiac electrical activity is represented on the electrocardiogram (ECG), although the depolarizations of the sinoatrial node, the atrioventricular node, and the His-Purkinje system involve too little tissue to be directly visible. In the ECG, the P wave represents atrial depolarization, the QRS complex signifies ventricular depolarization, and the T wave indicates ventricular repolarization. The PR interval, from the beginning of the P wave to the beginning of the QRS complex, measures the time it takes for the impulse to pass from the atrium to the ventricle. A large part of this interval reflects the slow conduction of the impulse through the atrioventricular node. The RR interval, measuring the time between consecutive QRS complexes, reflects the ventricular frequency. The QT interval, from the start of the QRS complex to the end of the T wave, indicates the duration of the ventricular depolarization. Normal QT interval values may vary slightly by gender and are influenced by the heart rate; therefore, the corrected QT interval (QTc) is often calculated. Cardiac rhythm disorders result from anomalies in the generation and/or conduction of the electrical impulse. Bradyarrhythmia occur when the intrinsic pacemaker system functions at a slower rate or when there are conduction blocks, in particular at the atrioventricular node or the His-Purkinje system. Most tachyarrhythmias result from a reentry mechanism, while others stem from increased normal automaticity or abnormal automaticity mechanisms.

1.1.4. MECHANISM OF TYPICAL REENTRY

Reentry describes the circular spread of an impulse along two interconnected pathways that have varying conduction speeds and different refractory periods. In this case, the atrioventricular node reentry is being used as an example. These two pathways connect the same points. Pathway A has a slower conduction velocity and a shorter refractory period, whereas Pathway B conducts normally but has a longer refractory period. As illustrated in Figure 3:

I. A normal impulse that arrives at point 1 travels along both pathways A and B. Conduction through pathway A occurs at a slower pace and encounters tissue at point 2, which is already depolarized and thus is in a refractory state. This scenario represents a typical sinus beat.

II. When a premature impulse encounters pathway B in a refractory state, it doesn't propagate through it. However, it can still be conducted along pathway A, which has a shorter refractory period. Upon reaching point 2, the impulse continues forward, but it can also retrogradely backtrack through pathway B. In here, it gets blocked by the refractory tissue at point 3. This situation corresponds to a premature beat (extrasystole) originating above the ventricles, resulting in an extended PR interval.

III. If the conduction through pathway A is slow enough, a premature impulse can retrogradely traverse the entire length of pathway B. This is possible because pathway B is now entirely out of its refractory period. If pathway A has also exited its refractory period, the impulse can re-enter pathway A and continue to circulate. This creates a scenario where, with each cycle, an anterograde impulse travel to the ventricle (4) and a retrograde impulse moves back to the atrium (5). This leads to sustained reentrant tachycardia.

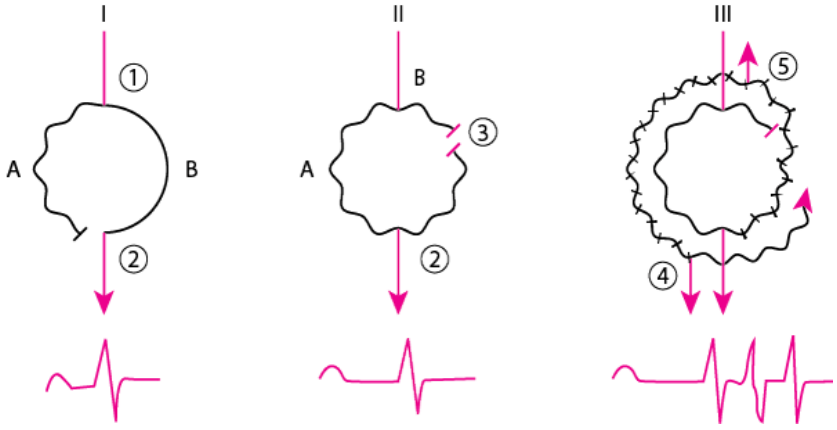

Figure 3. | *Image illustrating three specific situations: normal heartbeat, premature beat, and sustained reentrant tachycardia.*

Additionally, the abnormal P wave (P') and the delay in the atrioventricular node (prolonged P'R interval) are evident before the onset of tachycardia. Figure 4.

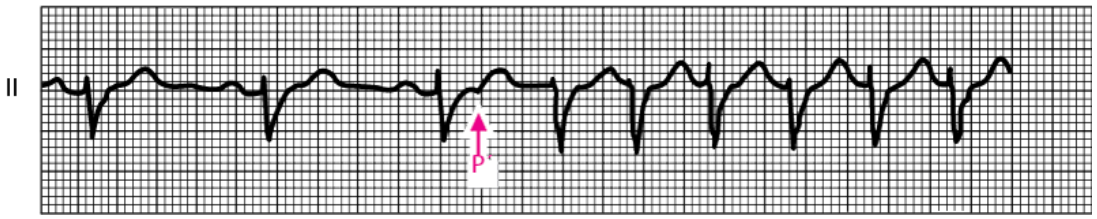

Figure 4. | *Image of an ECG trace at a point of tachycardia.*

In certain circumstances, typically following a premature beat (extrasystole), the reentry can result in a continuous circulation of the activation wave, leading to a tachyarrhythmia. Normally, the reentry is inhibited by refractory tissue during stimulation. However, three conditions can promote reentry:

- Reduction of tissue refractoriness
- Elongation of the conduction pathway
- Slowing of impulse conduction

From a diagnostic perspective, it is crucial not only to detect the anomaly in heart rate but, above all, to understand the localization within the cardiac muscle of the source point of the electrical issue.

1.2. CAUSES OF HEART FAILURE

Regarding the causes of heart failure, we can initially distinguish between two major groups of causes: ischemic and non-ischemic. In ischemic heart failure, the cause of left ventricular dysfunction can be traced back to coronary artery disease. Patients with this condition have often experienced one or more myocardial infarctions. The process following a large myocardial infarction often leads to progressive left ventricular dilation and the gradual onset of heart failure. It is also possible to have left ventricular dysfunction in the presence of severe coronary artery disease, even without clear clinical episodes of myocardial infarction. Among the causes of non-ischemic heart failure, the most common are related to valvular heart disease, long-standing hypertensive heart disease, and cardiomyopathy (both dilated and non-dilated), sometimes of genetic origin. Valvular diseases are a frequent cause of left heart failure and they can be divided into diseases of the aortic valve and diseases of the mitral valve, as well as diseases characterized by valve stenosis and those characterized by valve insufficiency. In general, when there is valve insufficiency, the heart undergoes what is commonly referred to as volume overload. Let's take aortic insufficiency as an example: Suppose that the left ventricle of an 80 kg individual ejects 100 milliliters of blood per beat at a rate of 60 beats per minute, resulting in a cardiac output of 6 liters per minute. After the ejection phase and the start of diastole, the aortic valve should be tightly closed to allow the blood ejected from the ventricle to move forward, perfusing the entire body. However, if the aortic valve doesn't close properly and allows blood to regurgitate into the ventricle, let's assume half of the systolic volume, i.e., 50 ml, is regurgitated. The effective output is then reduced to only 50 ml per beat, which is insufficient for the body, especially at a stable heart rate of 60 beats per minute. Additionally, the regurgitation of blood means that the left ventricle is overloaded with blood entering from both the aorta and the left atrium. To maintain adequate systolic output, the left ventricle undergoes eccentric hypertrophy (i.e., it tends to increase the thickness of the muscle outside the normal ventricular cavity), and as the condition worsens over the years, it progresses into a progressive ventricular dilation, a reduction in contractile force, and overt heart failure. [2] The other valvular alteration that leads to heart failure in the context of aortic valve disease is aortic stenosis. In this case the aortic valve is narrowed, meaning that the lumen that allows blood to be ejected from a ventricle is smaller than normal. This leads to an increased difficulty in ejecting blood from the ventricle and a higher need to generate greater systolic

pressure within the left ventricle to overcome the pressure difference (referred to as the gradient) between the left ventricle and the aorta. In this case, the response is concentric hypertrophy. In fact, the left ventricle responds to the need to generate more pressure by increasing the thickness of the muscular walls, much like how a weightlifting athlete's muscles grow in response to lifting increasing weights. In the initial stages of left ventricular hypertrophy, asymptomatic diastolic dysfunction is observed. However, as the condition progresses, it results in a typical scenario of heart failure with preserved ejection fraction. The left ventricle becomes much stiffer, and the end-diastolic pressure significantly increases. In the advanced stages of aortic stenosis (if valve replacement is not performed), the hypertrophy that has developed is no longer able to cope with the increased resistance to blood ejection, the muscle loses its contractile strength, and sometimes the ventricle dilates. Thus, it transitions from heart failure with preserved ejection fraction to heart failure with reduced ejection fraction. In these circumstances, aortic valve replacement becomes urgent, even at higher risk. [3] In the context of mitral valve diseases, mitral insufficiency causes volume overload of the left ventricle with an evolution similar to that described for aortic insufficiency; whereas mitral stenosis, by reducing the ease of blood access to the left ventricle, does not produce significant ventricular changes, but significantly increases the pressure in the left atrium and, therefore, an upstream in the pulmonary circulation, ultimately causing a dysfunction of the right heart as described earlier, and tricuspid valve insufficiency. Up until twenty or thirty years ago, when mitral stenosis was often caused by rheumatic disease in young individuals, this scenario was common and was referred to as the "tricuspidalization" of mitral stenosis. Another common cause of heart failure is hypertensive heart disease. Prolonged high blood pressure history leads the left ventricle to develop a higher pressure, and the disease's progression is similar to what was described for aortic stenosis. A more severe increase in ventricular thickness, entirely independent of high blood pressure, is seen in hypertrophic cardiomyopathy, a relatively common genetic disease that is important to recognize as it is associated with a risk of potentially fatal arrhythmias. Sometimes, among young athletes undergoing intense training, it is extremely challenging to distinguish whether the observed increase in left ventricular thickness is due to the so-called "athlete's heart" or if it represents the initial manifestation of hypertrophic cardiomyopathy. Because the latter possibility requires the suspension of sports activity and the initiation of a therapeutic approach, it is essential that this evaluation is carried out in centers with extensive experience. [4] Another cause of heart failure is non-ischemic dilated

cardiomyopathy. In this situation, the heart tends to progressively dilate, and its contractile strength gradually decreases, leading to a typical picture of heart failure with reduced ejection fraction. A subgroup of individuals with heart failure with reduced ejection fraction within non-ischemic dilated cardiomyopathy presents a familial predisposition to this disease, meaning a first-degree relative affected by the same condition [5].

1.3. CARDIOMYOPATHIES

A cardiomyopathy is a primary disease of the heart muscle. It distinguishes itself from other structural heart diseases such as coronary artery disease, valvular heart disease, and congenital heart diseases. Cardiomyopathies are primarily categorized into three main forms based on pathological features:

- Dilated
- Hypertrophic
- Restrictive

It is worth noting that the term "ischemic cardiomyopathy" refers to a condition that can affect patients with significant coronary artery disease, with or without infarcted areas, and is characterized by a dilated and hypocontractile ventricular myocardium. This category is not commonly included in the above-listed classifications because it does not describe a primary myocardial disorder. The clinical manifestations of cardiomyopathies typically align with those of heart failure and may vary depending on whether there is systolic dysfunction, diastolic dysfunction, or both. Some forms of cardiomyopathy can also lead to symptoms like chest pain, syncope, arrhythmias, or even sudden death. The evaluation generally involves a family history, blood tests, electrocardiography (ECG), chest X-ray, echocardiography, and cardiac magnetic resonance imaging. In some cases, endomyocardial biopsy could be necessary. If needed, further diagnostic investigations may be conducted to identify the cause of the cardiomyopathy. Treatment will depend on the specific type and underlying cause of the cardiomyopathy. [6, 7]

1.3.1. DILATED CARDIOMYOPATHY

As a primary myocardial dysfunction, myocardial dysfunction in dilated cardiomyopathy occurs in the absence of other disorders that could lead to myocardial dilation, such as severe occlusive coronary artery disease or conditions that impose pressure or volume overload on the ventricle (e.g., hypertension, valvular heart disease). In most patients, the anomaly affects both ventricles, in some cases, it affects only the left ventricle, and more rarely, only the right ventricle. As blood stasis becomes significant due to chamber dilation and dysfunction, mural thrombi can form. Tachycardic arrhythmias, as well as atrioventricular block, often complicate both acute myocarditis and late chronic dilatation phases. Atrial fibrillation typically arises when the left atrium has dilated.

1.3.2. HYPERTROPHIC CARDIOMYOPATHY

The myocardium exhibits alterations with cellular and myofibrillar disorganization; however, this finding is not specific to hypertrophic cardiomyopathy. In the most common phenotype, marked hypertrophy and thickening are observed in the anterior septum and in the anterior free wall contiguous below the aortic valve, with limited or absent hypertrophy in the posterior wall of the left ventricle. Isolated apical hypertrophy is sometimes seen, but virtually any type of asymmetric left ventricular hypertrophy may be observed, while symmetric hypertrophy is observed in a small minority of patients. Approximately 66% of patients exhibit an obstructive pattern, both at rest and during exercise. This obstruction results from mechanical obstacles of the left ventricular outflow during systole, caused by the anterior systolic motion of the mitral valve. During this anterior systolic motion, the mitral valve and the valvular apparatus are aspirated into the left ventricular outflow tract due to the Venturi effect, generated by high-velocity blood flow, leading to a flow obstruction and a reduced cardiac output. Additionally, mitral regurgitation may develop due to distortion of leaflet motion during anterior systolic motion of the mitral valve. These factors, namely, obstruction and valvular regurgitation, contribute to the onset of heart failure symptoms. Less frequently, hypertrophy of the midventricular region can result in an endocavitary gradient at the papillary muscle level, with rare risk of increased stress on the left ventricular wall and development of an apical aneurysm of the left ventricle. Hypertrophy results in increased stiffness and reduced compliance of the ventricular cavity (usually the left ventricle), hindering diastolic filling and leading to an increase in

telediastolic ventricular pressure and, consequently, in pulmonary venous pressure. This results in a reduced cardiac output, as filling resistance increases, especially when there is a gradient in the outflow tract. Tachycardia, by shortening filling time, tends to cause symptoms primarily during physical activity or in the presence of tachyarrhythmias.

1.3.3. RESTRICTIVE CARDIOMYOPATHY

Restrictive cardiomyopathy is a less common form of cardiomyopathy, which can be divided into two main categories:

• Non-obliterative: characterized by the abnormal infiltration of the myocardium by a foreign substance.
• Obliterative: characterized by endocardial and subendocardial fibrosis.

Both of these forms can occur in a diffuse or localized manner, affecting one or both ventricles, sometimes irregularly. When the myocardium thickens or becomes infiltrated, it can occur in one or both ventricles, usually the left ventricle. This can lead to malfunction of the tricuspid and mitral valves, resulting in valvular insufficiency. Furthermore, if nodal or conduction tissue is involved, the sinoatrial or atrioventricular node may function improperly, causing various degrees of sinoatrial and atrioventricular block. The primary hemodynamic consequence of this form of cardiomyopathy is diastolic dysfunction with a stiff, non-compliant ventricle and elevated filling pressures, which can lead to the development of pulmonary venous hypertension over time. Finally, if the compensatory hypertrophy of the infiltrated or fibrotic ventricles is inadequate to handle the workload, the systolic function may deteriorate. Additionally, in this condition, mural thrombi can form, posing a potential risk of systemic embolism.

1.4. INCIDENCE AND MORTALITY

According to the 2020 Istat data [8], in Italy, there were 63,952 deaths reported due to ischemic heart diseases in that year, with 34,095 being males and 29,857 females.
According to an article published by the Ministry of Health [9], diseases of the circulatory system caused 224,482 deaths (97,952 in men and 126,530 in women), accounting for 38.8%

of total deaths. Such a high percentage is partially attributed to the aging of the population and the low birth rates that have characterized the country in recent years. For ischemic heart diseases, there were 75,046 deaths (37,827 in men and 37,219 in women), accounting for approximately 33% of all deaths due to circulatory system diseases. In men, mortality is negligible until the age of 40, starts to emerge between 40 and 50, and then increases exponentially with age. In women, this phenomenon begins around the ages of 50-60 and increases rapidly. The disadvantage of men compared to women is more pronounced in the reproductive age and tends to decrease with advancing age. The difference in disease frequency between the two genders is also associated with differences in clinical manifestations, with sudden death and silent heart attacks being more frequent in women.

The term "incidence" refers to the number of new cases of a disease occurring in a population during a specific period, typically one year. Incidence data were derived from longitudinal studies conducted as part of the CUORE Project, which enrolled over 21,000 men and women aged 35-74 starting from the mid-1980s, with an average follow-up period of 13 years. The rates showed an incidence of coronary events (6.1 per 1,000 per year in men with a 28-day fatality rate of 28% and 1.6 per 1,000 per year in women with a 25% fatality rate). The fatality rate was 27.9% in men and 25.4% in women, increasing significantly with age. The data is presented in the Table 1.

Table 1. | *The project HEART reports a section of the table on incidence and fatality rates.*

Age (years)	**Coronary events**			
	Man		**Women**	
	Rates of incidence per year per 1,000	**Lethality, %**	**Rates of incidence per year per 1,000**	**Lethality, %**
35-44	3,2	9,6	0,5	8,3
45-54	4,5	15,3	1,2	11,4
55-64	9,7	33,6	2,8	27,1
65-74	10,1	54,2	4,5	54,5
35-74	**6,1**	**27,9**	**1,6**	**25,4**

1.5. TREATMENT

The treatment of arrhythmia-induced cardiomyopathy primarily focuses on managing the arrhythmia. This may involve its elimination or, alternatively, controlling the ventricular rate [10, 11], for example, in cases of persistent ventricular fibrillation. The therapeutic options mainly include beta-blockers, digitalis preparations, and amiodarone as suitable drugs for treating arrhythmia. Other antiarrhythmic agents may be considered only after a careful evaluation of the risk-benefit ratio. Therefore, the choice between these treatment approaches depends on the patient's age, pre-existing medical conditions, and the specific type of arrhythmia. In many cases, catheter ablation is the preferred long-term treatment. Additionally, in some instances, implantable cardiac electronic devices are used.

1.5.1. IMPLANTABLE CARDIAC DEFIBRILLATOR (ICD)

In recent years, there has been a growing interest in implantable cardiac electronic devices (ICDs), with a focus on refining both existing modes of cardiac pacing and defibrillation therapy and exploring new therapeutic strategies. The discussion regarding the effectiveness of ICD therapy in patients with non-ischemic cardiomyopathy is a topic of recent debate. Some studies [12] suggest that patients with right ventricular (RV) insufficiency who receive an ICD implant experience, improved survival compared to those without RV insufficiency but with baseline systolic dysfunction of the left ventricle. Another area of interest pertains to the subcutaneous defibrillator, which offers a less invasive and effective form of defibrillation therapy compared to traditional catheter-based devices. This technology has prompted investigations into maximizing its efficacy. Even in patients with a high body mass index, typically associated with lower success rates in subcutaneous ICD systems, positive outcomes have been achieved by optimizing the space between the device and the chest wall, positioning the pulse generator more posteriorly, and reducing shock impedance values. [13]

1.5.2. ABLATION

Transcatheter radiofrequency ablation has replaced antiarrhythmic drug therapy for the treatment of various cardiac arrhythmias. This approach offers several advantages, including symptom relief, improved functional capacity, enhanced quality of life, and the elimination of the need for long-term antiarrhythmic medication. It can also lead to long-term cost

savings. However, it is important to note that the procedure carries risks, which may vary depending on the specific ablation technique used and the experience of the healthcare provider performing the procedure. Therefore, it is crucial to carefully assess the risk-benefit ratio of radiofrequency ablation on an individual basis before proceeding with the intervention.

1.5.2.1. IMPORTANCE OF MAPPING

Indeed, before proceeding to the actual ablation of cardiac tissue, the procedure begins with the electrophysiological study of arrhythmia. This study involves the introduction of special diagnostic catheters into the heart chambers to electrically map the activity during atrial fibrillation, with the aim of understanding the origin and the extent of the issue. In addition to electrical mapping, a three-dimensional (3D) morphological mapping of the atria is performed. Before the procedure, various imaging options are available, including computed tomography, cardiac magnetic resonance imaging, and echocardiography. All data collected through these procedures are subsequently processed by specific programs with 3D reconstruction algorithms, which provide a more effective and real-time visualization of the heart. A clear and detailed representation of regional anatomy is essential to identify the region which to perform the ablation in. [14] Real-time visualization is also crucial for the catheter placement. In most ablation laboratories, the catheter manipulation is guided by fluoroscopy. However, fluoroscopy has the disadvantages associated with exposure to ionizing radiation. Often, cardiac mapping technologies and catheters are used in a complementary or even overlapping way to obtain maximum real-time information. There are advanced technologies that use a mapping catheter, which, through contact with the heart's walls, allows an anatomical reconstruction of the cardiac chambers. This technique is known as electroanatomical mapping (EAM). EAM systems also facilitate catheter navigation by enabling its three-dimensional localization in almost real-time, based on the simultaneous recording of spatial information and electrical activity from electrodes on the moving catheter. These systems combine anatomical information with electrophysiological data. Maps generated by these systems are extremely precise and allow the accurate identification of the areas that need to be treated through ablation, significantly reducing radiation exposure and radiation dose. [15]

1.5.2.2. PROCEDURE

Catheter ablation procedures are performed in an electrophysiology laboratory. [16] In this process, three or four electrode catheters are percutaneously inserted into a femoral, internal jugular, or subclavian vein and are positioned inside the heart to enable stimulation and recording at crucial sites. The effectiveness of trans-catheter ablation is closely tied to the accurate identification of the origin site of the arrhythmia. Once identified, an electrode catheter is placed directly in contact with the site, and radiofrequency energy is applied through the catheter to destroy it. Radiofrequency energy is delivered with wavelengths ranging from 300 to 750 kHz during trans-catheter ablation procedures. This process induces resistive heating of the tissue in contact with the electrode. Since the degree of tissue heating is inversely proportional to the radius to the fourth power, the lesions created by radiofrequency energy are small. Typical ablation catheters, with a diameter of 2.2 mm (7 French) and a 4 mm long distal electrode, create lesions of approximately 5-6 mm in diameter and 2-3 mm in depth. [17, 18] Larger lesions can be achieved with larger electrodes or with irrigated catheters using saline solution. Although electrical damage may contribute, the primary mechanism for tissue destruction by radiofrequency current is thermal damage. Irreversible tissue destruction requires the tissue temperature to reach about 50°C. In most ablation procedures, the power supplied by the radiofrequency generator is manually or automatically adjusted to maintain a temperature between 60 and 75°C at the electrode-tissue interface. [19, 20] If the temperature at the electrode-tissue interface exceeds 100°C, plasma clots and dried electrode tissue can form, hindering the effective flow of current, increasing the risk of thromboembolic complications, and necessitating the disposal of the catheter to allow the removal of the coagulated material from the electrode.

1.6. POTENTIAL PREDICTORS

In medicine, the term "Recurrence" refers to the reappearance of the symptoms of a disease in a patient who had previously been affected by it and had recovered. Recurrence is generally characterized by symptoms of a similar nature to the previous manifestations, but in other cases, it can present a more complex and severe clinical picture. To minimize this problem as much as possible, increasingly specific treatments have been developed. For example, in the SMASH-VT21 study, 128 patients with post-infarction cardiomyopathy and a history of ventricular tachycardia or ventricular fibrillation were randomized to therapy

with only an ICD (Implantable Cardioverter-Defibrillator) or therapy with ICD plus transcatheter ablation, the latter performed using anatomical substrate mapping in sinus rhythm. After a follow-up of approximately 2 years, 33% of patients randomized to ICD therapy experienced appropriate ICD interventions for arrhythmia recurrence, compared to 12% of patients treated with transcatheter ablation. [21] In the literature, there are many features that are considered to be correlated with recurrence, for example, Saglietto et al. [22] writes that the following pre-procedural, easily available, covariates were considered as potential candidate variables for the ML models training: age, gender, body mass index (BMI), estimated glomerular filtration rate (CKD-EPI formula were used), smoker status (active, former, never), hypertension, diabetes, dyslipidemia, history of heart failure, coronary artery disease, structural heart disease (valvular heart disease, dilated cardiomyopathy, hypertrophic cardiomyopathy), previous stroke/transient ischemic attack, presence of cardiac rhythm device (either pacemaker, implantable cardioverter defibrillator, or cardiac resynchronization therapy), hyperthyroidism, peripheral artery disease, chronic obstructive pulmonary disease, obstructive sleep apnea, CHA2DS2-VASc score, AF type (paroxysmal or persistent), history of atrial flutter, previous failed antiarrhythmic therapy, pre-procedural sinus rhythm, abnormal EKG (one or more of the following: atrioventricular block, bundle branch block, Q waves, ST-T abnormalities, and corrected QT > 460 ms), type of procedure (first ablation or re-do procedure), left ventricular ejection fraction (LVEF; %), left atrial (LA) anteroposterior diameter (mm), left ventricular end-diastolic volume (LVEDV; mL). While Croin et al. [23] writes that LVEF, the presence and extent of myocardial fibrosis evaluated through CMR-LGE, predict ventricular tachyarrhythmias in patients with ischemic and non-ischemic left ventricular dysfunction. The predictive value of LGE is independent of LVEF and whether the cardiomyopathy was of ischemic or non-ischemic etiology. Additionally, chronic obstructive pulmonary disease, age, general anesthesia, ischemic cardiomyopathy, New York Heart Association Class III or IV, ejection fraction, presentation with VT Storm, diabetes mellitus, and incessant VT are predictive factors. The non-inducibility of VT through PES after ablation is a predictor of VT recurrence.

2. OBJECTIVES

The issue with many types of interventions for critical and fatal diseases like these is the problem of recurrence. An intervention or treatment (pharmacological or otherwise) is performed, but clinicians don't know who will remain stable and who will experience the same symptoms again, or possibly even in a more severe form. For conditions like arrhythmias and heart diseases, extensive and in-depth studies have already been conducted. Pharmacological treatments and targeted interventions have been developed, especially in the case of ablation procedures. Furthermore, there are ongoing efforts to identify parameters that can predict prognosis. Prevention remains the best approach, but many parameters have been found, and there are numerous variables with uncertain correlations. In this situation of necessity, the techniques of analysis and prediction through Machine Learning are gaining traction. These new technologies can assist researchers in identifying parameters related to outcomes, and they can establish how these parameters are correlated with outcomes, often through various combinations. These techniques also hold promise for predictive purposes. The number of research studies and projects that combine medical knowledge with engineering is increasing every year. A single professional's role is no longer sufficient to address problems in this field. As a consequence, close collaboration between engineers and clinicians is increasingly necessary; this is the role of biomedical engineers. In recent years, as discussed previously, scientific articles on Artificial Intelligence applied to medical research have significantly increased. The aim of this thesis project is to offer engineering support for medical research finalized at addressing one of the most pressing and persistent problems. The number of analyzed patients is too low to establish scientific evidence, but a preliminary analysis has been conducted to develop a promising prototype to continue this research and increase the number of patients. The idea for this work is to use various Machine Learning techniques, such as Logistic Regression, Support Vector Machine, and Deep Learning, particularly Artificial Neural Networks, to discover which features are most correlated with outcomes and in what manner. This involves identifying which parameters from a patient's medical record can provide clinicians with insights into the likelihood of recurrence and how these parameters should be interpreted for this purpose. Furthermore, in this research, efforts have been made to integrate parameters obtained from electroanatomic mappings, not just for visual consultation, but to develop a dedicated algorithm capable of providing meaningful medical values to add to the patient's medical record and the database

used for analysis. The idea is also to compare the results obtained from the various methodologies used to provide a better and more accurate outcome, as well as to determine which of the proposed methodologies is best suited for this task.

3. MATERIALS AND METHODS

3.1. HARDWARE AND SOFTWARE DEVICES

For this study it has been used a computer with a i9 processor and a 64Gb RAM. As regards the software part, the program used was the MATLAB development environment in the R2023b version, both in the online and desktop versions. Within this development environment many toolboxes were used, including Statistic and Machine Learning Toolbox and Experiment Manager Toolbox for the analysis part, MATLAB App Designer for configuring the user interface, the MATLAB Compiler package for creating an app (all toolboxes inherent to this package are explained below). Moreover, parallel calculation has been implemented using the Parallel Computing Toolbox, to fully exploit all the processor cores, by performing calculations in parallel the risk of having cores that do not work is avoidable. The advantage is certainly reducing the calculation time during the execution of a program. In this project, the CARTO 3 System, designed and produced by Biosense Webster, was used for electroanatomical mapping.

3.2.ELECTROANATOMICAL MAPPING SYSTEM

Electroanatomic mapping, particularly when referring to the heart, is a process used to identify and chart the distribution in both time and space of electrical signals occurring during a specific cardiac rhythm. This technique is especially valuable during episodes of tachycardia, a rapid heart rate, as it helps understand how abnormal electrical activity develops, from when it starts to the point where the circuit closes. Furthermore, this process enables the identification of key target points for ablation, which involves the cauterization of specific cardiac areas to restore a normal heart rhythm. [24]

3.2.1. BEGINNING

The first non-fluoroscopic mapping system was introduced in clinical practice and marketed by a manufacturer founded by Shlomo Ben-Haim. In the following years, other companies introduced different systems, each with their unique features, sizes, and limitations, but all based on the same fundamental principles. [25] So all these systems, which have naturally evolved over time, share some fundamental principles. Each point sampled by a catheter and accepted as valid provides essential information, including its precise position in the microspace defined by the system, amplitude expressed in millivolts (mV), impedance, and activation time concerning a reference point in the cardiac cycle. The collection of sampled points is represented in three-dimensional space and in the temporal domain with static and dynamic maps of cardiac activation. In recent years, catheters designed solely for "high-resolution" mapping have been commercially developed, significantly altering the approach to interventional electrophysiology. While these catheters, combined with specific hardware and software, have varying construction characteristics, they share some design philosophies. The use of multiple electrodes allows a faster reconstruction of chambers and maps. Reduced electrode spacing and smaller electrode dimensions enhance spatial and temporal resolution of potentials. Finally, the flexibility of these catheter supports improves the adherence to the chamber walls. Below, in Figure 5, a comparison is shown between one of the early images obtained from these systems and an image from a high-resolution mapping system.

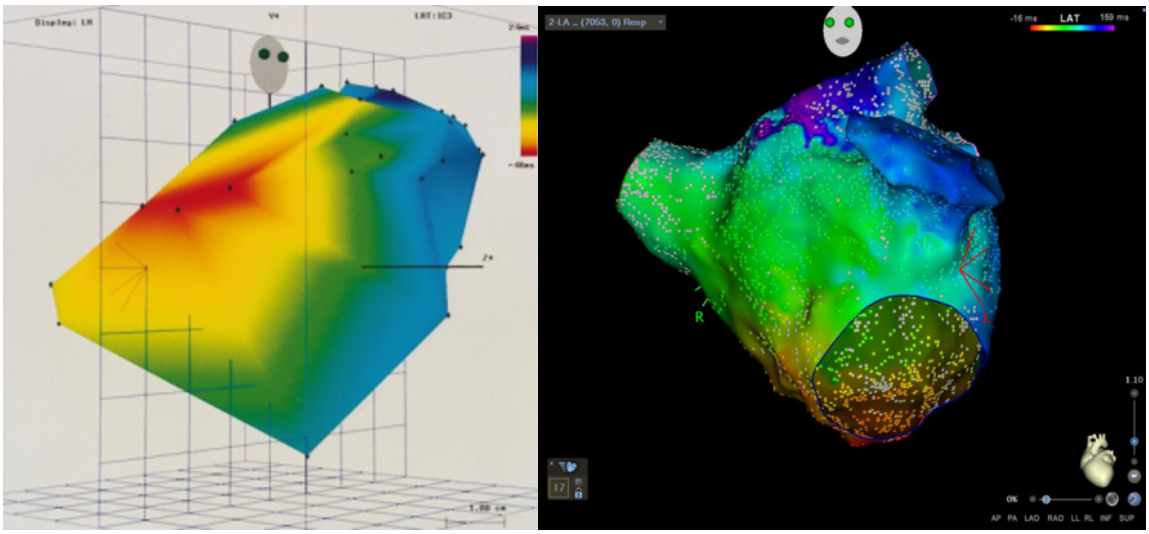

Figure 5. *| Comparison between two electroanatomic mapping images: on the left, one of the early images, and on the right, an image from one of the currently available high-definition systems.*

3.2.2. OPERATION

The Carto system consists of a low-intensity magnetic field generator composed of three coils positioned beneath the patient's chest Figure 6, six skin patches, three on the back and three on the patient's chest. Figure 7, a computer for data processing, and a display.

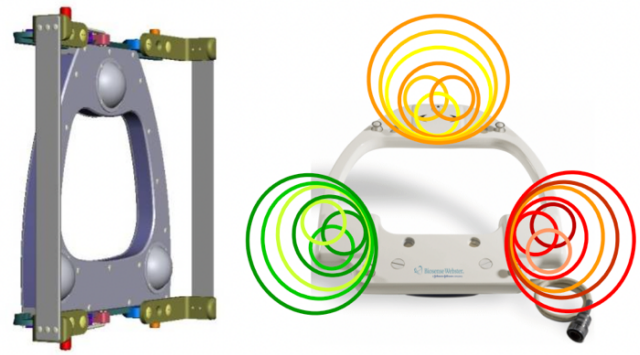

Figure 6. | *Two images related to the Location Pad of the CARTO 3 system.*

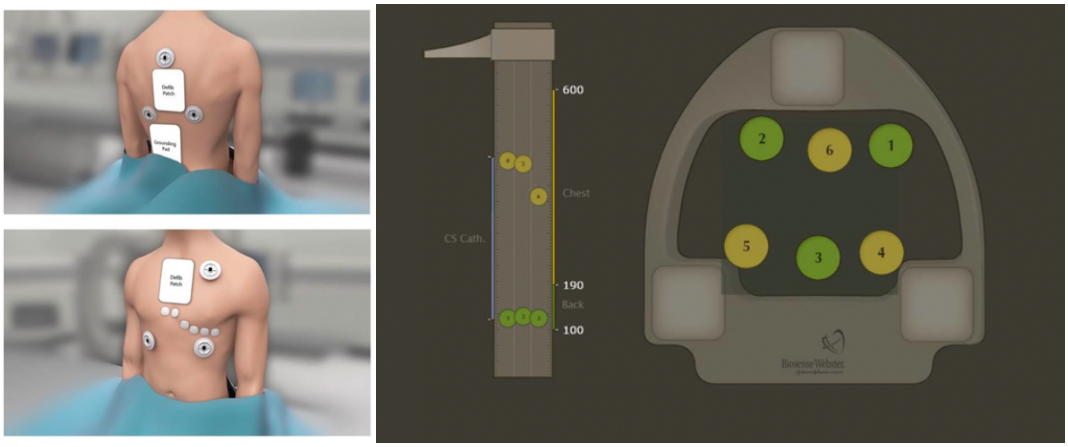

Figure 7. | *Left image shows the correct placement of patches on the patient's back and chest. The right image shows the localization of these patches within the reference system of the Location Pad.*

To perform 3D electroanatomic mapping of the cardiac chambers, specialized catheters with localization sensors in their tips are required. These sensors consist of spirals positioned orthogonally along the three spatial axes. The Carto system uses magnetic fields to determine the catheter's position and orientation and records intracavitary electrocardiograms from the sensors on the catheter's tip. By collecting spatial and electrical information from different points, the system reconstructs the real-time geometry of the cardiac chambers and analyzes arrhythmia mechanisms and ablation substrates. This process is based on the principle that metal spirals generate electric current when exposed to a magnetic field, with the current's

intensity depending on the strength of the magnetic field and the orientation of the spirals Figure 8.

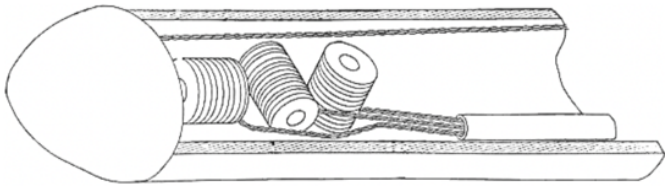

Figure 8. | *Image of the positioning and arrangement of spirals within the catheter used to generate a local reference system.*

The Carto system employs a triangulation algorithm similar to that used in GPS. The sensors on the catheter's tip measure the current intensity in each spiral (along the x, y, and z axes), allowing the system to determine the distance between the catheter and each magnetic field source. These distances are then used to create a spherical cap representing the possible position of the catheter towards each source. However, the catheter can only be located in the area where the spheres intersect, thus determining the three-dimensional position Figure 9.

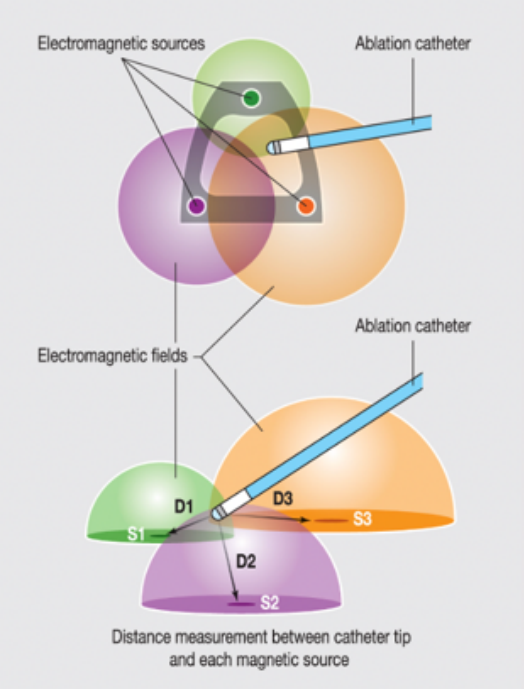

Figure 9. | *Superior and lateral images of the triangulation-based localization system.*

The Carto system can also calculate the catheter's roll, pitch, and yaw, in addition to the x, y, and z coordinates. Intracavitary electrocardiograms are recorded and integrated with

position information for each endocardial site reached, enabling the creation of the activation and cardiac geometry map.

To compensate for artifacts caused by heart and respiration movements, the Carto system makes corrections to the map coordinates, using the surface electrocardiogram as a reference and anatomical tags. The surface electrocardiogram is synchronized with the activation data recorded by the catheter during the map creation. The anatomical reference, which often is a skin patch or a catheter fixed inside the heart, is used to correct distortions due to the patient's thoracic movements Figure 10.

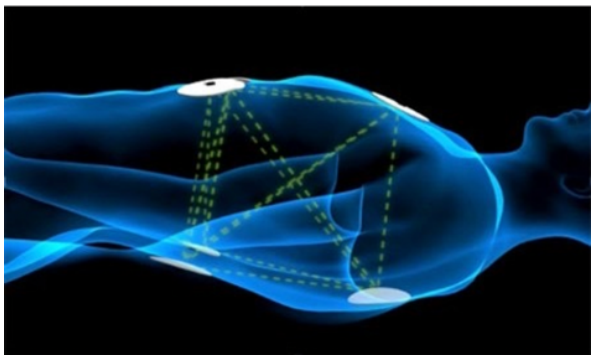

Figure 10. | *Explanatory image of the interaction between different patches placed on the patient to correct distortions due to breathing.*

Additionally, the system requires defining a "window of interest," representing the time interval, relative to a reference point on the surface electrocardiogram, in which local activation occurs, whether early or late compared to the reference. The total length of the window of interest cannot exceed the duration of the cardiac cycle in the case of tachycardia. The Carto system offers the option to overlay the electroanatomic map with CT or MR images acquired before the procedure, allowing verification of anatomical landmarks, improvement of cardiac geometry, and more precise guidance during ablation. In the latest version, Carto3, two additional modules are available: the CartoUNIVU module, which enables overlaying fluoroscopic images with the real-time electroanatomic map, and the CartoSound module, which uses intracardiac echocardiography as a tool for monitoring the procedure and providing anatomical support in map creation. [26, 27]

3.2.3. EXPORT

At the end of a measurement taken with this system, it is possible to download a file containing the results. This export includes various pieces of information for each contact point, meaning every point on the inner cardiac wall touched by the catheter. Among the many details we have, such as Point Index, catheter ID, the most relevant point-to-point values include:

• Position Coordinate, specifically three coordinates, one for each of the previously described imaginary axes. These are, of course, useful to determine the point's position.

• Angular Coordinate, meaning the three angles that determine the catheter's orientation concerning the fixed reference system at the signal acquisition moment. This specific information is not relevant to this project.

• Unipolar Voltage value uses a single electrode to record the electrical activity at a specific point within the heart. This type of recording measures the amplitude (intensity) of the electrical signal at that point.

• Bipolar Voltage value, on the other hand, uses two electrodes positioned at a certain distance from each other to record the electrical activity between them. This recording provides information about the direction and sequence of electrical impulse propagation between the two electrodes.

• Local Activation Time (LAT) value refers to the precise moment when cardiac cells in a specific point within the heart activate during the cardiac cycle. LAT represents the time elapsed from the beginning of the cardiac cycle or the onset of the electrocardiographic (ECG) wave, to when the cardiac cells in a particular region start contracting in response to the electrical impulse. In other words, LAT indicates when the electrical impulse reaches that specific point and begins triggering the contraction of cardiac cells in that area. This can help determine whether a point or an entire region is delayed or early in its activation.

• Impedance value is expressed in Ohms and refers to the voltage difference between two points in the electrical circuit (i.e., between the electrodes). This voltage difference is influenced by the resistance of cardiac tissues. The higher the impedance, the greater the electrical resistance offered by the tissues. This is because when an electrical impulse spreads through the heart tissues, it encounters a certain resistance, represented by impedance. This resistance affects the shape and amplitude of the electrical signals recorded by the electrodes. Impedance is primarily used to assess the quality of contact between the electrodes or catheters and the heart tissues. A significant change in impedance could indicate a contact issue or a less-than-ideal electrode position, which could affect the accuracy of electrical activity measurements. The export is downloaded as a digital file, available in various formats, with the most used being text files (.txt) and Excel files (.xlsx).

3.2.4. EXTRACTION ALGORITHM

This type of analysis, in addition to the image of the endocardial cavities, provides a type of point-to-point information, as described above. This means that all the values that can be calculated or estimated are referred to the single point selected each time by the catheter.
In this study the parameters that we want to calculate, starting from this analysis, are:

• Evaluation of the extension and dispersion of late potentials.
• Point-to-point difference of the bipolar potential with unipolar potential with possible identification of more organized regions, in which multiple points with high differences are grouped in the same area.
• Areas of deceleration, this happens when there is a very early and very late potential in a short range, i.e. with a large difference in the timing of activation.

In order to comply with this request, an ad hoc algorithm was created.
The main problem with these mapping systems is that of the entire image displayed on the monitor only some points are measured, other values in between are estimated and calculated by the system. The created algorithm works only with the values of the points measured for a more accurate analysis. Then there is the problem that each patient's heart is of different dimensions and positioned with a slight difference in orientation compared to another, it was therefore decided to normalize these values by projecting them onto a sphere. More

precisely, a sphere of fixed dimensions is constructed, the entire surface of the sphere is divided precisely into 14400 faces of equal area, the centroid of the point cloud resulting from the measurements carried out is calculated, the centroid is positioned at the center of the sphere, and all points are projected from the centroid to the inner surface of this sphere. Each face of the surface of the sphere can have only two values: null value if the face has not been hit by any projection, otherwise the face takes on the values corresponding to the projected point. The number of faces into which to divide the sphere was chosen following many tests and that value was selected for which no more than one projected point corresponded to each face, this in order not to reduce the accuracy of the analysis and not have to make any kind of approximation. The major advantages of this projection are two: one referring precisely to the position problem described above, and the other one to the ability to calculate the required values. The first one is because the coordinates of each face are known a priori and that the dimensions of both the sphere and the faces are fixed, and therefore are the same for each measurement. This allows us to compare all patient mappings with each other, which would not have been possible without normalization in terms of space and orientation. The second one is because without an algorithm like this the export would be difficult to understand and then unusable. In the export many values are reported for each point, but being able to interpret the distances and differences between the values simply by reading them is too difficult and not at all pleasant. Instead, thanks to this algorithm we can automatically extract not only the individual point-to-point values but also values, as in the case of the required parameters, which are calculated as iterations between values of different points. In particular, for the assessment of the extent and dispersion of late potentials, the LAT value is used for all projections. Values below a certain threshold (last 20% between the maximum and minimum values) are considered late, and adjacent late points are part of the same delay area. This allows the calculation of the extension and position of late potentials as a percentage relative to non-late ones. Following the same criteria, positions, and extensions of areas with early potentials are calculated. This helps identify deceleration areas, particularly their value. By applying a similar process, amplitudes of point-to-point potentials, both bipolar and unipolar, are calculated. The difference is calculated for each individual point, and those with values above a certain threshold (last 20% between the maximum and minimum values) are considered critical. This way, it is possible to assess the presence of regions where multiple points with significant differences, i.e., critical values, are clustered in the same area. From this, we

calculate the exemption as a percentage of the total points and their position. Below, in Figure 11, two images resulting from an execution of the designed program are shown. In addition to useful numerical results for the continuation of this project, it was thought that the possibility of visually seeing what is being carried out would also be useful for the clinician. Methods for providing the clinician with this algorithm intuitively without having to have knowledge of the MATLAB development environment, or other prior programming knowledge, are explained below.

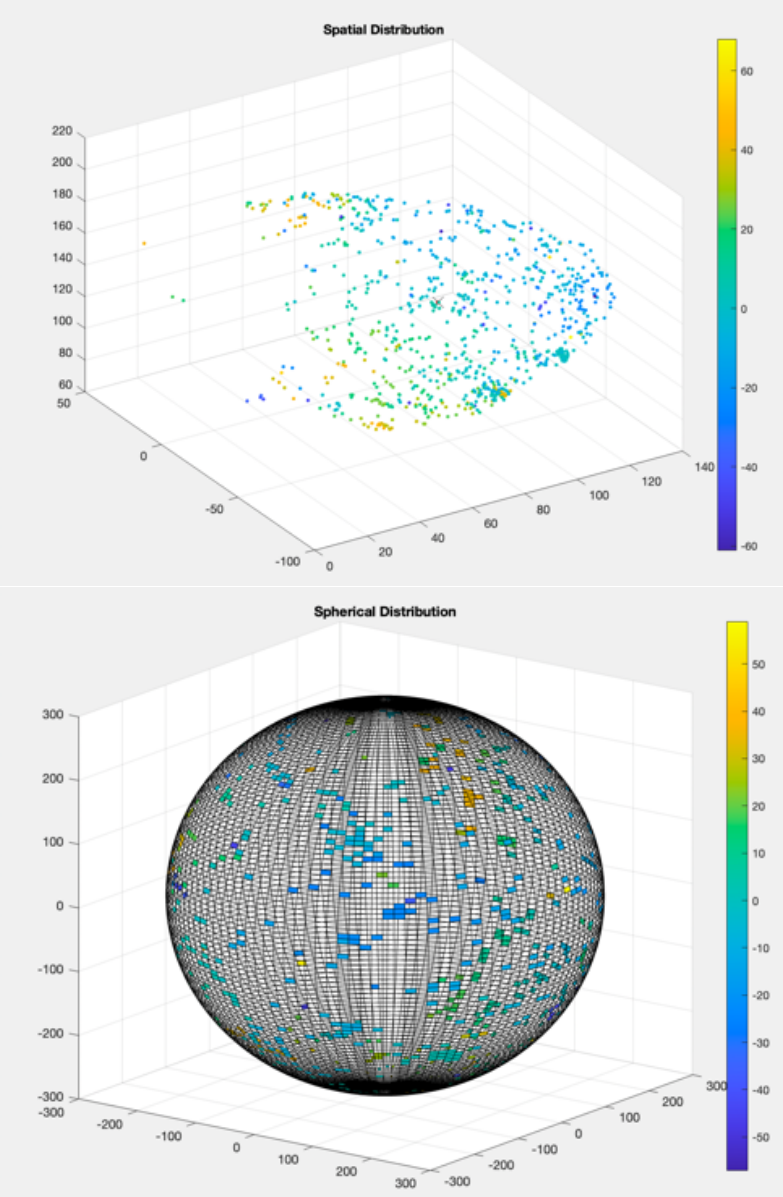

Figure 11. *| Two images resulting from the extraction algorithm. The upper image depicts the spatial distribution of points in the left ventricle measured during an examination with the CARTO 3 system. The lower image represents the spherical distribution of the same points.*

3.2.5. APPLICATION DIFFUSION

In the MATLAB development environment, there is MATLAB App Designer, a toolbox for creating graphical user interfaces. Once a project (.prj) is created, in the "Design View" section, you can build the graphical part of the interface, defining window dimensions, colors, or adding buttons, progress bars, and other components from the component library. In the "Code View" section, you need to insert the program within functions for the selected components. Once the interface is complete, you can download it as an application using specific toolboxes based on the type of disclosure you want to follow. The types of distribution are divided into two groups: sharing with other MATLAB users and sharing with non-users.

3.2.5.1. FOR MATLAB USERS

If you want to share an application with a MATLAB user, you can do that by downloading your interface as a toolbox (.mltbx) or as an executable (.mlapp) that can be shared with other users. In this case, there is no need to create or install any additional packages because the application will use MATLAB to function. If the user does not have MATLAB installed on their computer, it is still possible to use the online version and run the application via a web browser. The advantages are that the created application takes up very little space, is easy to share, and does not encounter any compatibility or executability issues. The only drawback is that the end-user must be a MATLAB user. An intuitive diagram is proposed below in Figure 12.

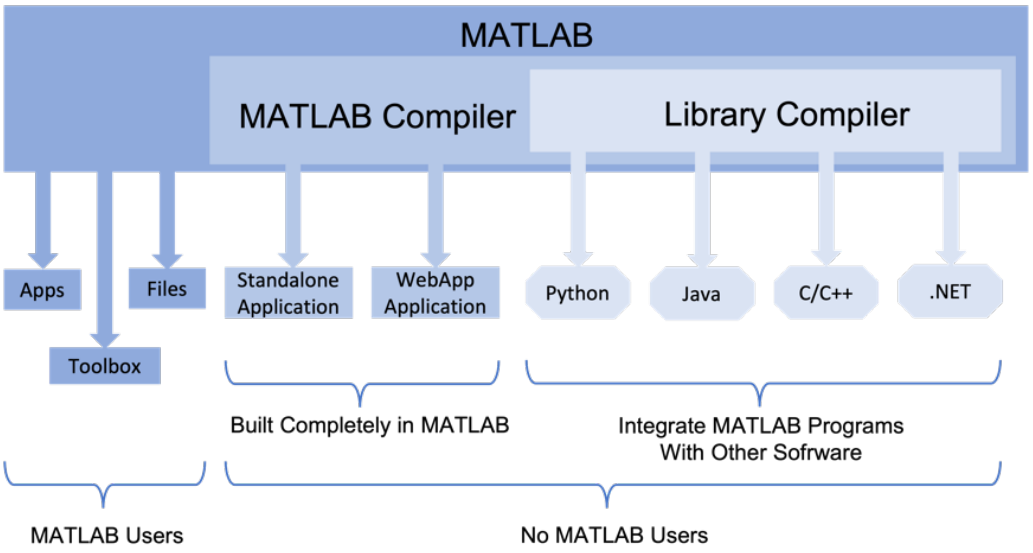

Figure 12. | *Block diagram summarizing the diffusion methods of an application designed in MATLAB.*

3.2.5.2. DESKTOP APPLICATION

In case you want to create a version of the application for use on a computer where MATLAB is not installed or for someone who is not a user, you can use the Application Compiler toolbox, which creates a standalone application from the user interface you've designed. This solution allows you to distribute the application to those who do not have MATLAB, but they must still download MATLAB Runtime, which contains a set of libraries that enable its execution. MATLAB Runtime can be downloaded from the website: https://www.mathworks.com/products/compiler/matlab-runtime.html, or it can be added as a package during the application installation. The advantage is the ability to create an application independent of the MATLAB development environment. The disadvantages are compatibility and updates. The first one because creating a Windows executable requires a computer with a Windows operating system, and the same applies to macOS for Mac and Linux. The second one because any updates to the application require installing the new version, and the old one cannot be modified.

3.2.5.3. ONLINE VERSION

If you want to share the application online, you can use MATLAB Web App Server. To do this, create the application as you would for other methods using MATLAB App Designer and the MATLAB Web App Compiler toolbox. You create a package that will be hosted on

the MATLAB Web App Server, which can be configured locally or in the cloud. In the local option, you can host the apps on a server within the local network and allow access to the applications to all users via the generated URL without the need to install MATLAB or other components. The main drawback is that the computer running MATLAB Web App Server acts as a server for users, so to keep the application active, the PC must remain on at all times. MATLAB Web App Server is executed on your local computer, so when you turn off your computer, the server is interrupted, and the web page with the application becomes inaccessible. Alternatively, you can deploy MATLAB Web App Server in the cloud using sites such as Amazon Web Services, Microsoft Azure, or Google Cloud Platform. Each of these offers paid packages that allow application sharing. For example, to use Amazon Web Services, you need a VPN, which is provided by Amazon for a fee. You also need an EC2 instance to host the application, and the instance costs vary based on storage space. For added security, you must use an encrypted key pair (.pem) to access the instance. You must then download a copy of MATLAB onto the instance, transfer the files, and run the application. Once you have tested and configured security on the instance, you can share the IP address or URL of the configured web app server. There is a free trial period for 12 months for Amazon Web Services, but the instances are still paid. The only two free instances have low performance and limited memory. The VPN is only paid for specific zones; most available zones are free. Much of the procedure is explained in the "help" sections of MATLAB and Amazon, but not all steps are well described.

3.2.5.4. LIBRARY FOR OTHER SOFTWARE

Alternatively, you can convert the .mat file into a function and use the Library Compiler toolbox to create a library compatible with other programming languages like C, C++, Python, or Java. This way, you can create the web page or application using another programming language. This is an advantage for those who are more familiar with other languages, especially for desktop applications. For online cases, the advantage is using a programming language compatible with most servers. The significant drawback is that by reducing the entire script to a function, you can execute limited functions, and not all operations that could be done in another way are compatible.

3.2.5.5. PROJECT SHARING

First, we tried to follow a free and user-friendly procedure for an end user who does not know a development environment like MATLAB. First, we created a new trial account on MATLAB because with an institutional account, no procedures could be followed to upload your application online. Once the server was set up, we had to access our license on MathWorks and authorize the computer's Host ID hosting the server. However, this is not possible if the license is not owned. Afterward, the initial idea was to use MATLAB Web App Server locally. However, the issue was not having a computer dedicated exclusively for this purpose. We then decided to upload it to the cloud and chose one of Amazon's servers. We used all the available free options. For the VPN, there are many choices, and for the instance, we chose between the two available ones. We then selected the operating system and the update to be downloaded onto the instance, created a key pair for access, and performed a test. The connection was very slow because the servers are far from Italy or Europe, and the VPN had limited performance. The instance was the second problem because the specifications were low. In addition to having only a few cores to rely on, the memory was also insufficient, with a portion being used to host the operating system, and the remaining part was not able to host MATLAB. The same procedure was followed with Microsoft Azure and Google Cloud Platform, but the same problems were encountered in each attempt. A desktop application was created, but this could only be run on a computer with the same operating system as the source computer. Therefore, we thought that the most effective solution was not free. After comparing prices, the most cost-effective and practical solution was to create a MathWorks account for the end user and upload the application to the "Drive" folder of the same account. A link is then provided to access MATLAB Online, and once logged in, you can view and run the application in the "Current Folder" section.

3.3. STUDY COHORT

We retrospectively collected data from consecutive patients referred to the Clinical Cardiology and Arithmology Department of the University Hospital of the Ospedali Riuniti di Ancona from September 2018 to July 2023. The patients for this study are selected with several inclusion and exclusion criteria. The first inclusion criterion is certainly that of the presence of some cardiac disease, ischemic or non-ischemic, in the medical record. Moreover, of all the patients in the department, all those who had carried out electroanatomic mapping, in addition to the classic clinical tests which we will discuss in depth later, were selected for this study. To be included in the study cohort the patient must have undergone follow-up, otherwise it would be impossible to predict the outcome. Regarding the exclusion criteria, the fundamental thing was relevance; of the entire medical record, the features that according to the literature are not indicative for the prediction of heart disease were neglected and therefore not included in the database. Those features with an excessive number of missing values, i.e., and those parameters not common to all patients, were excluded. The features with non-heterogeneous binary logical values were also eliminated, this is because it is useless to use a parameter that always contains the same value as input for our model, they would have been parameters without any statistical significance.

3.4. CLINICAL TESTS

All the patients present in the database, in addition to the medical history conducted by the doctor, conducted some clinical tests from which specific values considered risk factors for this field of study were extracted. The clinical tests in question are: echocardiogram, electrocardiogram, magnetic resonance imaging, electroanatomic mapping and blood tests. Obviously, follow-up was then conducted for each patient to verify the patients who relapsed and those who did not.

3.5. DATA PREPARATION

The input values for our models, regardless of the machine learning technique used, are always the same. The inputs are numerical or binary values, obviously depending on the type of variable, and refer to one or more specific parameters emerged from the anamnesis, follow-up, or clinical tests. Numerical values are expressed as a percentage or in a specific

unit of muscle, while the binary values 0 and 1 often represent the logical "yes/no" value which most often indicates whether that parameter is present in the patient or not. Binary values also express belonging to two different classes, such as males and females in the "sex" parameter. When there are more than two classes then the variable represents a value ranging from one to the number of classes to indicate its membership. According to the literature, there are various parameters that can be indicative of some type of heart disease, which is why only a few values considered valid for the study conducted were chosen from the entire patient medical records. Specifically, the values in question refer to: risk factors and pathological conditions, symptoms, characteristics of heart disease, echocardiogram, electrocardiogram, magnetic resonance imaging, electroanatomic mapping, reason for the procedure, ablation surgery, blood tests. The "recurrence" parameter of the database is instead selected as the expected value, wanting to precisely predict this. The value we want as the output of our models is a logical one, where the value 1 corresponds to that patient who has relapsed and the value 0 to the patient who has not. Specifically, 223 features were initially provided for each patient. Following a careful analysis and having applied the inclusion and exclusion criteria, 162 features were discarded and 61 were therefore taken into consideration for the analysis. One feature, namely "Recurrence" is the variable to be predicted. The other 60 used for the study are summarized in Table 2. The "Variables" column shows the names or acronyms of the 60 features used for this study. For a better understanding, all acronyms or abbreviations are spelled out in full, are listed below, in Table 3. The type of variable is indicated in the "Type" column. The letters V, B and C correspond respectively to: Value, to indicate whether the variable takes on real values, Binary, where the values 0 and 1 generally imply the presence or absence of that parameter, and Class, in the event that the variable takes on values for a certain range. In the last three columns there are the average values for type V variables, the count of the presence of the parameter in the case of type B variables and the number of the most present class in the case of type C variables. In "Total" they are reported these values for all 220 patients, in "No" the values referring to patients who do not relapse and in "Yes" those referring to patients who relapse.

Table 2. *| The variables used for the study are reported, indicating the type of variable: N for numerical value, B for binary, C for categorical. The average values or the count of the presence or absence of the variable are also provided for all subjects and for those who have experienced recurrence and those who have not.*

Variables	**Type**	**Total (n=220)**	**Recurrence**	
			No (n=175)	**Yes (n=45)**
Age (years)	V	55	54	62
Sex	B	M=180/F=40	M=137/F=38	M=43/F=2
BMI	V	0,0026	0,0026	0,0028
Hypertension	B	87	62	25
Diabetes mellitus	B	15	10	5
Smoking	B	69	49	20
Family history of MCI	B	14	10	4
OSAS	B	10	7	3
BPCO	B	8	5	3
Vascular disease	B	34	22	12
Prior TIA/STROKE	B	14	9	5
Previous angioplasty	B	38	28	10
Bypass surgery	B	13	7	6
FA	B	35	21	14
HF	B	121	80	41
NYHA Class	C	0	0	2
HFpEF	B	62	44	18
HFmEF	B	16	16	0
HFrEF	B	59	34	25
COVID19	B	25	22	3
Anemia	B	7	5	2
Palpitations	B	67	55	12
Dyspnea	B	41	32	9
Lightheadedness	B	23	16	7
Syncope	B	13	9	4
Chest pain	B	28	22	6
Fatigue	B	16	9	7
Dilated cardiomyopathy	B	41	30	11
Ischemic cardiomyopathy	B	46	30	16
Myocarditis	B	20	15	5
Valvular cardiomyopathy	B	24	19	5
VT Idiopathic	B	61	59	2
FE (%)	V	49,12	50,98	41,91
LAV (ml/m2)	V	33,71	32,66	39,04
RVD (mm)	V	37,48	37,40	37,92
TAPSE (mm)	V	22,48	22,69	21,62
Mitral valve insufficiency	B	174	137	37
Tricuspid valve insufficiency	B	160	130	30
Aortic valve insufficiency	B	54	43	11
PAPs (mmHg)	V	28,55	27,87	31,12
LV aneurysm	B	22	15	7
Rhythm	B	39	24	15
T-wave inversion	B	36	33	3
LVEDV (ml/m2)	V	92,58	92,40	93,80
LGE	B	107	97	10
BEV	B	118	112	6
Arrhythmic storm	B	49	23	26
TV Paroxysmal	B	74	57	17

Ablation (Yes/No)	B	155	113	42
Presence of late potentials	B	94	61	33
Substrate ablation	B	139	102	37
Bipolar endocardial low-voltage area	B	76	55	21
Bipolar endocardial scar area	B	59	39	20
Inducibility (yes/no)	B	13	9	4
HB (g/dl)	V	13,90	13,97	13,63
RDW	V	13,37	13,24	13,88
Blood glucose	V	98,87	93,82	118,69
Creatinine	V	1,01	0,97	1,16
Percentage uni/bi potential (%)	V	0,314	0,242	0,547
Percentage LAT area (%)	V	11,400	11,701	10,423
Gradient value (ms)	V	9,238	8,842	10,523

Table 3. | *The full meaning of the acronyms or abbreviations used to indicate the considered variables is provided. In cases where the word 'or' is used between two acronyms, it is because some variables are presented in the Italian language.*

BMI	Body Mass Index
MCI	Unexpected Cardiac Death
OSAS	Obstructive Sleep Apnea Syndrome
BPCO or COPD	Chronic Obstructive Pulmonary Disease
Prior TIA/STROKE	Preceding Transient Ischemic Attack or Stroke
FA or AF	Atrial Fibrillation
HF	Heart Failure
Classe NYHA	New York Heart Association functional classification
HFpEF	Heart Failure with Preserved Ejection Fraction
HFmEF	Heart Failure with Mid-Range Ejection Fraction
HFrEF	Heart Failure with Mid-Range Ejection Fraction
FE o EF (%)	Ejection Fraction
LAV (ml/m2)	Left Atrial Volume
RVD (mm)	Right Ventricular Diameter
TAPSE (mm)	Tricuspid Annular Plane Systolic Excursion
PAPs (mmHg)	Pulmonary Artery Pressure
LV	refered to the Left Ventricol
LVEDV (ml/m2)	Left Ventricular End-Diastolic Volume
LGE	Late Gadolinium Enhancement
BEV	Ventricular Ectopic Beats
TV or VT	Ventricular Tachycardia
HB (g/dl)	concentration of Hemoglobin in the Blood
RDW	Red cell Distribution Width
Percentage uni/bi potential (%)	expansion of areas with high potential difference Unipolar and Bipolar
Percentage LAT area (%)	extension of Areas with delay
Gradient value (ms)	Gradient Value

3.6. TRAINING, VALIDATION AND TESTING

The database used is made up of 220 patients, the division into training sets and test sets was made with a percentage of 80% for training and 20% for testing. Having few samples available, we chose to use K-fold cross-validation to evaluate the model. For the purpose of validation, it is standard practice to consciously partition the data into training and testing sets. The data in the testing set is not utilized for model training, allowing us to mimic real-world scenarios where the model encounters previously unseen variables. Cross-validation is a widely employed technique that extends this approach. It involves repeatedly dividing the data into distinct training and testing (validation) sets, known as "folds," with different combinations. This iterative process can be performed as many times as needed to generate an averaged assessment of model performance. By doing so, it reduces the risk of overfitting the model to the data, thereby enhancing the model's ability to make predictions that generalize well to a broader population. [28] The number of folds is 3. The parameters that were used to evaluate the performance of the various models are: AUC, Sensibility, Specificity, Accuracy and Precision. The Area Under the Receiver Operating Characteristic Curve (AUC) is a metric employed in binary classification tasks. The Receiver Operating Characteristic (ROC) curve plots the true-positive rate against the false-positive rate, and the AUC quantifies the portion of the curve area. An AUC of 0.5 corresponds to random classification, while an AUC of 1.0 signifies a model that makes flawless predictions. Sensitivity, also known as Recall or True Positive Rate, represents the ability of a classification model to correctly identify all positive cases present in the test data. A high sensitivity indicates that the model is effective in detecting the presence of the class of interest, minimizing false negatives. Sensitivity = TP / (TP + FN). Specificity measures a model's ability to correctly identify negative cases. It expresses the percentage of true negatives compared to the total negative cases and indicates how accurately the model is able to distinguish examples that do not belong to the class of interest. Specificity = TN / (TN + FP). Precision is a measure of the fraction of positive instances correctly identified by the model out of the total number of instances identified as positive. It is calculated as the ratio of the number of true positives to the sum of true positives and false positives. Accuracy focuses on the accuracy of positive predictions. Precision = TP / (TP + FP). Accuracy, on the other hand, is a general measure of the overall correctness of the model. Represents the fraction of all correct predictions out of the total number of predictions. Accuracy evaluates

the overall correctness of the model, taking into account both positive and negative predictions. Accuracy = (TP + TN) / (TP + TN + FP + FN). [29]

3.7. MACHINE LEARNING METHODS

Artificial intelligence (AI) has promised to revolutionize medicine for over 30 years, and there have been technological breakthroughs in recent years that could make this a reality, including exponential increases in computing power, big-data processing technologies, access to large clinical data sets using electronic health records, and machine learning (ML). [30] In the field of medicine, ML has the potential to improve the accuracy of diagnostic algorithms and personalize patient treatment. The fundamental concept of ML is to employ algorithms that take in input data, apply computer analysis to predict output values within an acceptable range of accuracy, discern patterns and trends within the data, and ultimately learn from experience. While ML is not a new concept and has been around since the advent of modern computing, the idea of a thinking machine has been proposed to harness the computational capacity of computers to uncover patterns and draw conclusions that may be challenging to attain through conventional statistical methods. These traditional methods often depend on human operators to formulate and provide a rule base or assumptions regarding correlations for further computer analysis. [31] ML is either founded upon or incorporates statistical foundations to underpin its functioning. [32]

3.7.1. REGRESSION

Linear regression is arguably the simplest ML algorithm. The central concept in regression analysis is to establish a connection between one or more numeric features and a single numeric target. Linear regression is an analytical method employed to address regression problems by employing a straight line to characterize a dataset. In the case of univariate linear regression, which focuses on predicting a target value using just a single feature, it can be represented in a slope-intercept form:

$$Y = \beta_0 + \beta_1 X$$

In this representation, β_1 serves as the slope weight, describing how much the line rises on the y-axis for each increment in x. The intercept, β_0, indicates the point where the line

intersects the y-axis. [33, 34] Linear regression models a dataset using this slope-intercept form, with the machine's task being to ascertain values of a and b that enable the determined line to best correlate the provided x values with the y values. To be more precise:

$$\beta_1 = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} \quad \text{and} \quad \beta_0 = \bar{Y} - \beta_1\bar{X}$$

Where X, Y are the detected values and $\bar{X}, \bar{Y}$ are the theoretical values. Multiple linear regression is similar; however, there are multiple weights in the algorithm, each describing to what degree each feature influences the target. Basically, there is rarely a single function that fits a dataset perfectly. To measure the error associated with a fit, the residuals are measured. Conceptually, residuals are the vertical distances between predicted values, $\bar{Y}$, and actual values, Y. For multiple linear regression the model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Logistic regression is a classification algorithm where the goal is to find a relationship between features and the probability of a particular outcome. Instead of employing the straight line generated by linear regression to estimate class probability, logistic regression uses a sigmoidal curve to estimate class probability. This curve is determined by the sigmoid function:

$$y = \frac{1}{1 + e^{-x}}$$

which produces an S-shaped curve that transforms discrete or continuous numeric features (x) into a single numerical value (y) between 0 and 1. The key advantage of this approach is that probabilities are bounded within the range of 0 and 1 (i.e., probabilities cannot be negative or exceed 1). Logistic regression can be either binomial, where there are only two possible outcomes, or multinomial, where there can be three or more possible outcomes. [33, 34] In statistics, the logistic model (or logit model) is a statistical model that shapes the probability of an event taking place by expressing the log-odds for the event as a linear combination of one or more independent variables. In regression analysis, logistic regression (or logit regression) entails estimating the parameters of a logistic model, which are the coefficients in the linear combination. Formally, in binary logistic regression, there is a single binary dependent variable, coded using an indicator variable, where the two values are labeled "0" and "1," while the independent variables can each be a binary variable (two

classes, coded by an indicator variable) or a continuous variable (any real value). The logistic function is therefore represented as follows:

$$logit(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Notice that the right hand side of the equation above looks like the multiple linear regression equation. However, the technique to estimate the regression coefficients in a logistic regression model is different from that used to estimate the regression coefficients in a multiple linear regression model. In logistic regression the coefficients derived from the model (e.g., β_1) indicate the change in the expected log odds relative to a one unit change in X_1, holding all other predictors constant. Defined p as the probability, the multiple logistic regression model can be written as follows:

$$p = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n}}$$

p is the expected probability that the outcome is present; X_1 through X_n are distinct independent variables; and β_1 through β_n are the regression coefficients. In this thesis work, the first machine learning method was chosen to use a model based on multinomial logistic regression. The choice of a logistic regression was made because the variable to be predicted is binary, so the outputs we expect can only be 0 or 1. Multinomial because the parameters used as predictors are multiple. In MATLAB, in the function used for the regression, it was specified that the model was linear and that the distribution was binomial. Several models equal to all the combinations of five variables were proposed, for each combination the test performance was calculated to see which was the best.

3.7.2. SUPPORT VECTOR MACHINE (SVM)

Support Vector Machines (SVMs) are one of the cornerstones of machine learning and are particularly powerful for binary classification. In this chapter, we will delve into the theory behind SVMs, including the details of how to find the optimal hyperplane and the optimization problem. Furthermore, we will explore the use of nonlinear kernels, including the polynomial and sigmoid kernels. Support Vector Machines (SVMs) are a machine learning model used for both classification and regression tasks. The primary goal of an SVM is to find an optimal hyperplane in a multidimensional space that can effectively

separate different data classes. This hyperplane is chosen to maximize the margin between classes, which is the distance between the hyperplane and the nearest data points from each class, referred to as "support vectors". The equation of a hyperplane in an N-dimensional space is given by:

$$x^T\beta + \beta_0 = 0$$

Where:

• β is a weight vector that determines the orientation of the hyperplane.
• x is an input vector.
• β_0 is the bias term that regulates the position of the hyperplane relative to the origin.

In a Figure 13, the left panel shows the separable case. The decision boundary is the solid line, while broken lines bound the shaded maximal margin of width 2M = 2/||β||. The right panel shows the non-separable (overlap) case. [33, 34] The points labeled ξ^*_{1j} are on the wrong side of their margin by an amount ξ^*_j = Mξ_j; points on the correct side have ξ^*_1 = 0. The margin is maximized subject to a total budget $\xi_i \leq$ constant. Hence, ξ^*_j is the total distance of points on the wrong side of their margin.

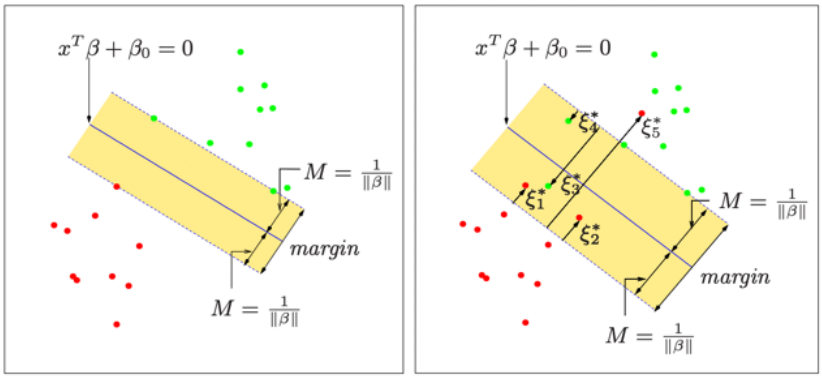

Figure 13. | *Explanatory image of the theory behind the Support Vector Machine method in a linear case.*

Consider a p-dimensional real-valued space (e.g., $\mathbb{R}p$). An optimal separating hyperplane is essentially an p-1 dimensional affine space residing within the larger p-dimensional space. For p=2, this affine space is simply a one-dimensional line, while for p=3, it is a two-dimensional plane. For higher dimensions, this affine space is known as a hyperplane. This

is certainly challenging (if not impossible) to visualize, but it can be conceptually grasped. Note that "affine" refers to a hyperplane that doesn't necessarily pass through the origin (or the zero element) of the larger space. If we consider elements in the p-dimensional space, that is, $x = (X_1, \ldots, X_2) \in \mathbb{R}p$, such an affine hyperplane p-1 dimensional is defined by the following equation:

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = 0$$

or equivalently:

$$\beta_0 + \sum_{j=1}^{p} \beta_j X_j = 0$$

We can construct a maximum-margin hyperplane (MMH), which is the separation hyperplane that is farthest from any training observations. First, you compute the perpendicular distance from each training observation x_i to a given separation hyperplane. The closest perpendicular distance from a training observation to the hyperplane is known as the margin. MMH is the separation hyperplane where the margin is the largest. This ensures that it is the farthest minimum distance from any training observation. The classification procedure is then simply a matter of determining which side a test observation falls on. Such a classifier is known as a maximum-margin classifier (MMC). We hope that a wide margin on the training observations also leads to a wide margin on test observations and therefore provides a good classification rate. However, note that we must be cautious to avoid overfitting when the number of feature dimensions is high. In this case, overfitting means that the MMH fits the training data very well but can perform quite poorly when exposed to test data. One of the key features of MMC (and subsequently of SVM) is that the position of the MMH depends solely on the support vectors, which are the training observations that lie directly on the margin boundary, but not on the hyperplane. See points A, B, and C in Figure 14.

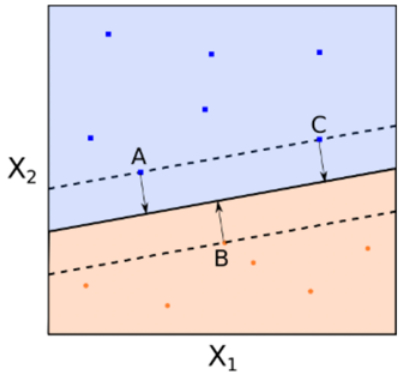

Figure 14. | *Image related to explaining the search for the best hyperplane.*

This means that the position of the MMH does NOT depend on other training observations. The MMH is the solution to the following optimization procedure:

$$\max_{\beta, \beta_0} M$$

$$subject\ to \quad y_i(x_i^T \beta + \beta_0) \geq M, \quad i = 1, \ldots, N$$

Where:

• β is a weight vector that determines the orientation of the hyperplane.
• β_0 is the bias term that regulates the position of the hyperplane relative to the origin.
• (x_i, y_i) are the training points, with x_i representing the input vector and y_i the class label (+1 or -1).

In many cases, the data is not linearly separable in the original feature space. This is where nonlinear kernel functions come into play. A kernel function is a transformation that maps the data into a higher-dimensional space where linear separation is possible. An example in Figure 15.

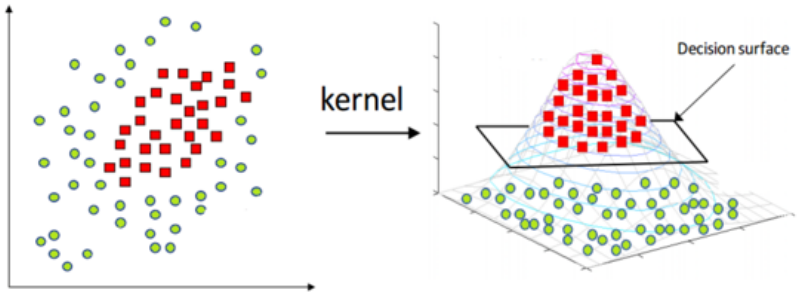

Figure 15. | *Explanatory image on the utility of using nonlinear kernel functions.*

Common examples of nonlinear kernel functions include the Radial Basis Function (RBF) kernel, the polynomial kernel, and the sigmoid kernel. [33, 34] The Radial Basis Function (RBF) kernel is defined as:

$$K(x_1, x_2) = \exp\left(-\frac{||x_1 - x_2||^2}{2\sigma^2}\right)$$

Where $||x_1 - x_2||$ is the Euclidean distance between points x_1 and x_2, and σ is a scale parameter.

The polynomial kernel is defined as:

$$K(x_1, x_2) = (x_1^T x_2 + c)^\rho$$

Where ρ is a positive integer representing the degree of the polynomial, and c is a constant.

The sigmoid kernel is defined as:

$$K(x_1, x_2) = tanh(\beta_0 x_1^T x_2 + \beta_1)$$

Where β_0 and β_1 are parameters that regulate the shape of the sigmoid function.

The analysis using Support Vector Machine was conducted with specific MATLAB functions. Specifically, the classification problem has been solved, the output variable to be predicted is always of a binary logical type, hence with only the values 0 and 1. The "ClassName" is defined and it is used to name the output classes, this helps the algorithm to understand that it is a classification and by defining this parameter as “[False True]” it helps the algorithm to understand that the variable to be predicted is binary, and that therefore the outputs it should expect will be “False ” in the case of patients who do not relapse and “True”

in the case of patients who relapse. Another parameter defined was the "BoxConstraint", it is a positive number that determines the "rigidity" of the margin. Larger “BoxConstraint” values correspond to tighter margins, making the model more sensitive to classification errors on training data. Conversely, smaller values of “BoxConstraint“ lead to wider margins, making the model more tolerant to misclassification on training data, but potentially at the expense of lower generalization ability. The “Solver” parameter is another parameter that has been defined, this specifies the algorithm used to solve the optimization problem associated with training the SVM. The solver determines how the optimization problem underlying the search for support vectors and associated weights is solved. The two main values are: “SMO” (Sequential Minimal Optimization), this is the default algorithm. It is based on minimal sequential optimization and is particularly effective for moderately sized problems. The minimal sequential approach divides the optimization problem into smaller subproblems, iteratively optimizing the weights associated with pairs of training examples. The second is “ISDA” (Iterative Single Data Algorithm), this solver focuses on a single training example at a time. It is useful when working with very large datasets where storing the complete kernel matrix might be prohibitive. The “KernelFunction” parameter specifies which type of function you want to use, while the “KernelScale” parameter optimizes the predictors for the “KernelFunction” specification. First of all, a linear analysis was done using Support Vector Machine, the "ClassName" was specified, the "KernelFunction" was defined as "Linear" to indicate a linear analysis, "BoxConstraint" was set to 100 to strongly penalize the classification errors, trying to obtain a separation hyperplane that separates the classes more rigorously. The "Solver" has been left with the default value, this is because the chosen variables are only used 5 at a time as predictors so we have no need for optimization algorithms that take space into account. The second analysis done with Support Vector Machine was not linear. The parameters have all remained as indicated before except the "KernelFunction" which was defined as "rbf" because the radial basis function was chosen as the kernel. This is a great choice when it comes to capturing complex, non-linear relationships in your data. This kernel is flexible and can adapt to a wide range of data distributions. Then the "KernelScale" parameter with the value "Auto" was also added because we wanted to leave the algorithm the possibility of inserting the most suitable value for the type of dataset provided.

3.7.3. ARTIFICIAL NEURAL NETWORK (ANN)

The Artificial Neural Networks (ANNs) are computational models inspired by the structure and functioning of the human brain. They have become a key tool in machine learning and data processing. In this chapter, we will explore how ANNs work, their fundamental components, and how they are trained for classification and regression tasks. ANNs are comprised of a series of interconnected layers of artificial "neurons" or nodes. The basic structure of a neural network includes the following layers Figure 16:

• Input Layer: This layer accepts raw input data and passes it to the rest of the network.

• Hidden Layers: These intermediate layers, which can be one or more, perform complex computations to learn intermediate representations of the data.

• Output Layer: This layer produces the final results of the network.

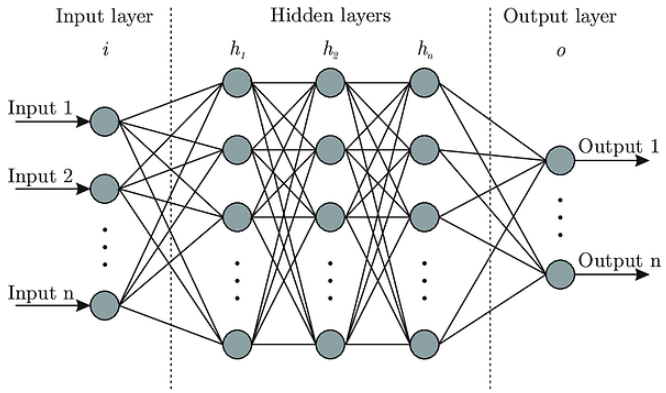

Figure 16. *| Image describing the classical architecture of an Artificial Neural Network and its division into layers composed of different interconnected neurons.*

Each connection between nodes has an associated weight that models the importance of each connection. The functioning of an artificial neuron is similar to that of a biological neuron. An artificial neuron receives input from previous neurons, computes a weighted sum of the inputs, applies an activation function, and produces an output. The formula for a neuron can be expressed as follows:

$$y = f\left(\sum_{i=1}^{n}(w_i x_i) + b\right)$$

Where:

• y is the output of the neuron.
• $f()$ is the activation function.
• x_i represents the inputs from the previous neuron.
• $w_i$are the weights associated with the inputs.
• b is the bias term.

More precisely, [33, 34] the Input x_i: Every artificial neuron receives input from one or more previous neurons or input data. These inputs are multiplied by the associated weights w_i. The weights represent the importance of each input in the neuron's computation. The weighted sum: the products of the weights w_i and the inputs x_i are summed. This weighted sum represents the level of activation of the neuron, i.e., how "activated" the neuron is based on the inputs. The Bias term b: It is an additional parameter that influences the neuron's activation. This term allows shifting the neuron's output upward or downward. Basically, the bias helps better model the neuron's behavior. Lastly, the Activation Function $f()$: the weighted sum (previous output) is then processed through an activation function. The activation function introduces nonlinearity into the neuron's output. Common activation functions include the sigmoid function, the Rectified Linear Unit (ReLU) function, the hyperbolic tangent (tanh) function, among others, Figure 17.

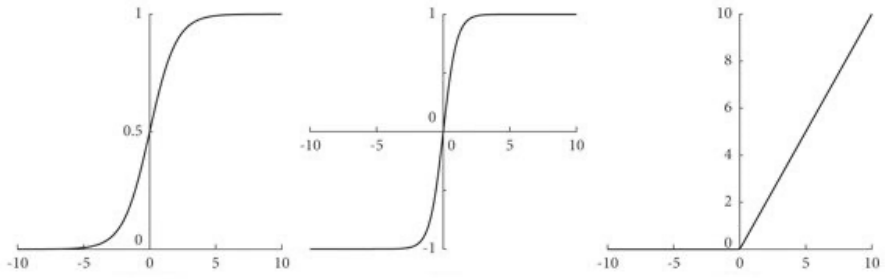

Figure 17. *| Three graphs related to activation functions for Artificial Neural Networks. On the left, a Sigmoid function; in the center, a Hyperbolic Tangent function; and on the right, a Rectified Linear Unit (Relu) function.*

Activation functions are crucial components of ANNs as they introduce nonlinearity into the model. The mentioned activation functions are as follows:

• Sigmoid:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

• ReLU (Rectified Linear Unit):

$$f(z) = \max(0, z)$$

• Hyperbolic Tangent (tanh):

$$\tanh(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$

Training a neural network involves optimizing the weights and biases so that the network can make accurate predictions on the data. The most common training algorithm is backpropagation with gradient-based optimization, where the weights are updated iteratively to minimize a cost function. The cost function measures the error between the network's predictions and the actual values. A common cost function for classification is "cross-entropy," while for regression, mean squared error (MSE) is often used. Artificial Neural Networks (ANNs) are known for their ability to learn from complex and nonlinear data. This makes them suitable for a wide range of applications, including image recognition, natural language processing, financial forecasting, and more. Their adaptability is another strong point, as they can be used to solve classification, regression, and clustering problems. A distinctive advantage of ANNs is their ability to detect complex patterns in data. They can uncover nonlinear relationships and capture subtle details, making them effective in situations where other machine learning techniques might fail. Additionally, ANNs exhibit a remarkable tolerance to noise in data and a good generalization ability, meaning they can make accurate predictions on new data. ANNs can be executed in parallel on specialized hardware, speeding up the training and evaluation process. This aspect contributes to their efficiency in computationally intensive applications. However, there are also disadvantages associated with the use of Artificial Neural Networks. Firstly, they require significant computational resources, including computing power and memory. Training large neural networks on extensive datasets can demand expensive hardware. A common issue associated with ANNs is overfitting, which occurs when the network fits too closely to the training data and does not generalize well to test data. To mitigate overfitting, regularization techniques are often necessary. Furthermore, ANNs can lack interpretability. Understanding how a neural network makes decisions can be elusive, which can be problematic in applications requiring transparency. Another challenge is the need for a large amount of training data to

achieve high performance. In some situations, there may not be enough data available to train an effective network. Training complex neural networks can be time-consuming, especially on large datasets. It's also important to note that ANNs may struggle to handle minority classes in imbalanced classification problems if the training dataset has few examples for those classes. Finally, the choice of parameters, such as the number of layers, the number of nodes, and the activation function, may require a lot of experimentation and optimization. In general, Artificial Neural Networks are a powerful and flexible tool, but it's essential to carefully consider the specific advantages and disadvantages for the problem at hand before deciding to use them as a solution. Specifical functions in MATLAB were also used in this case. The output variable is always of the same type. Here the parameters that have been specified are the "LayerSize" which specifies how many layers there are and how many neurons are contained in each layer; and the “Activation” parameter which indicates which activation function you want to use. Experiment Manager Toolbox is a MATLAB tool that was used to establish what was the best combination of parameter values to adjust. The functionality of the toolbox is very simple and intuitive, you choose the input table, i.e. the database you want to provide. You specify which of the variables present is the variable to predict and select the number of predictors. In the case of this study, 5 predictors were selected at random, the optimization analysis of the parameters was carried out, a new combination of 5 predictors was chosen and the optimization analysis was conducted again. At the end of the processing, a model was chosen from among those with the highest performance for our dataset. The parameters were saved and inserted into the classification function. In detail: the “LayerSize” is “[10 10 10] and the “ActivationFunction” chosen was “Relu”.

4. RESULTS

The results obtained, which follow, include the performance values of the various models for each Machine Learning technique adopted. These values are compared, the presence or absence of features, the recurrence in the various models and in the different methods used are then also compared and discussed. For each model, we then analyze and discuss how the features referring to that model affect the predictability of the output, i.e., how much they are correlated with it and whether they are negatively or positively correlated. To do this, the parameters associated with the various models are analyzed. In the case of linear

classifications, you will have a formula, this is for the case of Logistics Regression for both the Linear SVM, consisting of an intercept value and five coefficients associated with the five featurs taken each time and indicating their weight with respect to the output. Therefore, in the case of linear models, the interpretability of the results is guaranteed, i.e., every detail of the results obtained is known. In the case of non-linear models, however, we choose to observe the non-linearities at the expense of interpretability. Hence, it is not easy to make sense of what has been achieved. In the case of the non-linear SVM, since it was not possible to assign a weight directly to the individual variables, SHAP Values were used. SHAP, an acronym for SHapley Additive exPlanations, is a theoretical approach based on game theory and is often used in the interpretation of Machine Learning models. The goal of SHAP is to fairly attribute the contribution of each feature to a model's prediction, in order to obtain more understandable and transparent explanations. SHAP can be used to explain model behavior on a global scale, providing a general understanding of the relationships between features and model output, and on a local scale, specifically explaining the prediction for a particular observation. Shapley values indicate how much each feature contributes to the difference between the model's prediction and its average prediction. They can therefore be used to interpret the results even though they are not coefficients; in fact, the coefficients of a linear SVM are directly associated with the features, allowing a clearer interpretation. Each coefficient represents the effect of the respective feature on the model's decision. A nonlinear SVM with Local Shapley Values, on the other hand, can capture complex nonlinear relationships between features and model output. Thus, we chose to explore both options and evaluate which one provides a better understanding of the model with respect to the specific requirements of the problem.

4.1. LINEAR

4.1.1. LOGISTIC REGRESSION

The best six logistic regression models were reported and analyzed. The following table, Table 4, shows the variables assigned to each model. This allows us to identify the most present parameters and give an idea of the frequency with which they are present.

Table 4. | *The most recurrent variables in the six selected models among the best results of logistic regressions are reported, with "X" indicating the presence of the variable in that model.*

VARIABLE	MODEL 1	MODEL 2	MODEL 3	MODEL 4	MODEL 5	MODEL 6
NYHA Class	X	X	X	X	X	X
PAPs	X	X	X	X	X	X
Rhythm	X					
Percentage LAT area					X	
Gradient value	X	X	X	X	X	X
HFrEF				X		
Hypertension	X					
Sex		X				X
Creatinine		X				
HFmEF			X			
Arrhythmic storm			X	X	X	X

The performance values of the six selected models are shown in Table 5.

Table 5. | *The performance parameters of the models from the analyses conducted with logistic linear regression are reported. The performances are expressed in terms of AUC, Accuracy, Precision, Sensibility, and Specificity.*

	AUC Train	AUC Test	Accuracy	Precision	Sensibility	Specificity
MODEL 1	0,997	0,975	0,881	0,700	0,537	0,960
MODEL 2	0,980	0,934	0,875	0,696	0,532	0,951
MODEL 3	0,991	0,925	0,867	0,691	0,529	0,948
MODEL 4	0,986	0,925	0,848	0,687	0,527	0,945
MODEL 5	0,986	0,925	0,837	0,685	0,527	0,941
MODEL 6	0,986	0,925	0,833	0,685	0,525	0,940

Table 6, however, shows the parameters of the equation resulting from the logistic regression analysis. In the first line we find the value of intercept, while in the other lines the variables for each model are listed. Each column corresponds to a model, and the coefficients of that specific model relating to that specific variable are inserted in the intersection between rows and columns. This is used to realize the influence of the variable on the output. The various weights can be compared within the same model, and it can be verified in which direction the variable affects the output.

Table 6. | *The coefficients of the polynomial resulting from logistic linear regression analyses are reported. In the "Variable" column, features are listed, and the values of each feature are inserted in the column of the respective model.*

VARIABLE	COEFFICIENT					
	MODEL 1	MODEL 2	MODEL 3	MODEL 4	MODEL 5	MODEL 6
Intercept	-21,007	-22,114	-12,052	-26,637	-12,457	-21,434
NYHA Class	3,889	5,161	1,281	5,060	2,697	2,548
PAPs	0,262	0,279	0,144	0,313	0,124	0,383
Rhythm	-0,944					
Percentage LAT area					0,043	
Gradient value	0,492	0,682	0,375	0,441	0,271	0,223
HFrEF				2,104		
Hypertension	2,638					
Sex		-98,661				-104,115
Creatinine		-1,133				
HFmEF			-98,818			
Arrhythmic storm			2,808	0,208	1,223	3,089

4.1.2. LINEAR SVM

To study the presence and frequency of features in the linear SVM model, a table, Table 7, similar to the one used in the regression results, is reported. Also, in this case the best six models were selected.

Table 7. | *The most recurrent variables in the six selected models among the best results of linear support vector machines are reported, with "X" indicating the presence of the variable in that model.*

VARIABLE	MODEL 1	MODEL 2	MODEL 3	MODEL 4	MODEL 5	MODEL 6
FE	X					
TAPSE	X	X	X			
PAPs	X	X	X	X	X	X
Arrhythmic storm	X	X	X	X	X	X
Gradient value	X	X	X	X	X	X
Rhythm		X				
T-eave inversion			X	X		
TV Paroxysmal				X		
Creatinine					X	X
Percentage uni/bi potential					X	
Percentage LAT area						X

Also in this case, the performance values of the selected models have been reported, Table 8

Table 8. | *The performance parameters of the models from the analyses conducted with logistic linear regression are reported. The performances are expressed in terms of AUC, Accuracy, Precision, Sensibility, and Specificity.*

	AUC Train	AUC Test	Accuracy	Precision	Sensibility	Specificity
MODEL 1	0,996	0,992	0,884	0,708	0,538	0,961
MODEL 2	0,993	0,988	0,871	0,701	0,536	0,956
MODEL 3	0,993	0,988	0,867	0,692	0,535	0,952
MODEL 4	0,984	0,979	0,851	0,655	0,532	0,949
MODEL 5	0,984	0,979	0,821	0,648	0,529	0,942
MODEL 6	0,984	0,979	0,817	0,619	0,528	0,940

In order to interpret the models, i.e., the equations relating to each model, Table 9 shows the intercepts and coefficients of each equation relating to the various models.

Table 9. | *The coefficients of the polynomial resulting from linear analyses of support vector machines are reported. In the "Variable" column, features are listed, and the values of each feature are inserted in the column of the respective model.*

VARIABLE	COEFFICIENT					
	MODEL 1	MODEL 2	MODEL 3	MODEL 4	MODEL 5	MODEL 6
Intercept	-8,301	-3,474	-1,096	-2,207	-9,353	-1,096
FE	-0,046					
TAPSE	0,119	0,04	0,076			
PAPs	1,963	0,234	0,287	0,179	0,297	0,335
Arrhythmic storm	1,963	3,635	4,217	4,906	4,751	3,971
Gradient value	0,155	0,148	0,338	0,175	0,288	0,111
Rhythm		1,818				
T-eave inversion			2,573	1,769		
TV Paroxysmal				0,408		
Creatinine					-1,268	-2,408
Percentage uni/bi potential					0,328	
Percentage LAT area						0,007

4.2. NON-LINEAR

4.2.1. KERNEL SVM

The best six models were chosen for the analysis carried out with non-linear SVM. Table 10 shows the variables associated with the chosen models.

Table 10. | *The most recurrent variables in the six selected models among the best results of support vector machines with radial basis function (rbf) kernel are reported, with "X" indicating the presence of the variable in that model.*

VARIABLE	MODEL 1	MODEL 2	MODEL 3	MODEL 4	MODEL 5	MODEL 6
Bipolar endocardial scar area	X	X	X	X	X	X
Gradient value	X	X	X	X	X	X
VT Idiopathic	X	X	X	X	X	X
LV aneurysm		X				
Rhythm						X
FA	X	X				
BPCO	X		X	X		X
HFmEF					X	
BMI			X			
Diabetes mellitus				X		
Family history of MCI					X	

In Table 11 you can find the performance results of the models chosen for the non-linear SVM.

Table 11. | *11 The performance parameters of the models from the analyses conducted with logistic linear regression are reported. The performances are expressed in terms of AUC, Accuracy, Precision, Sensibility, and Specificity.*

	AUC Train	AUC Test	Accuracy	Precision	Sensibility	Specificity
MODEL 1	0,995	0,986	0,889	0,724	0,537	0,969
MODEL 2	0,994	0,984	0,878	0,711	0,535	0,961
MODEL 3	0,986	0,967	0,867	0,692	0,534	0,953
MODEL 4	0,983	0,967	0,854	0,675	0,533	0,949
MODEL 5	0,967	0,945	0,832	0,644	0,530	0,947
MODEL 6	0,954	0,945	0,822	0,629	0,528	0,944

For the case of non-linear models, it is not possible to have coefficients directly referring to the variables, we therefore chose to use the SHAP values. More specifically, the average of the absolute value of the SHAP Values is reported for each model. In Figure 18 you can see how individual variables influence the output in that specific model.

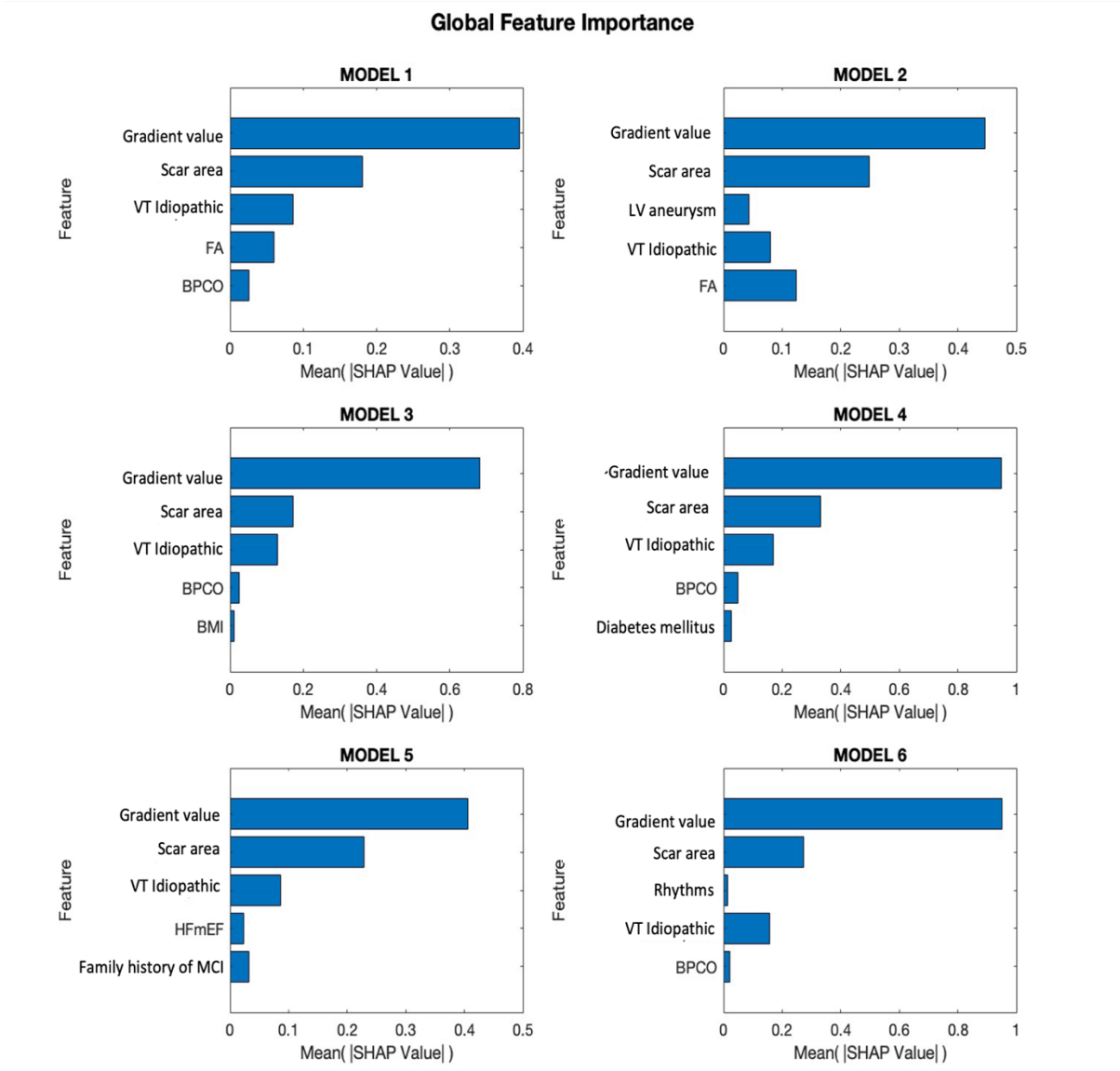

Figure 18. *| For each model of support vector machine with rbf kernel function, the SHAP values of the features are reported. Specifically, the mean of the absolute value.*

4.2.2. ARTIFICIAL NEURAL NETWORK

As the last method used, we implemented Artificial Neural Networks, built with input layers of 10 connections, a hidden layer of 10 and an output layer of 10 connections, and using the "relu" activation function. Table 12 shows the best six models built with 5 variables.

Table 12. | *The most recurrent variables in the six selected models among the best results of artificial neural networks are reported, with "X" indicating the presence of the variable in that model.*

VARIABLE	MODEL 1	MODEL 2	MODEL 3	MODEL 4	MODEL 5	MODEL 6
Sex	X	X	X	X	X	
Diabetes mellitus	X			X		
NYHA Class	X	X	X		X	X
VT Idiopathic	X	X	X	X		
Gradient value	X	X	X	X	X	X
BPCO		X				
FA			X			
HFpEF				X		
PAPs					X	X
Rhythm					X	
BMI						X
TAPSE						X

Table 13 shows the performances obtained from the chosen models.

Table 13. | *The performance parameters of the models from the analyses conducted with logistic linear regression are reported. The performances are expressed in terms of AUC, Accuracy, Precision, Sensibility, and Specificity.*

	AUC Train	AUC Test	Accuracy	Precision	Sensibility	Specificity
MODEL 1	0,993	0,986	0,887	0,710	0,537	0,965
MODEL 2	0,990	0,982	0,874	0,702	0,535	0,963
MODEL 3	0,986	0,978	0,861	0,686	0,533	0,956
MODEL 4	0,986	0,978	0,843	0,662	0,531	0,952
MODEL 5	0,984	0,974	0,817	0,631	0,530	0,950
MODEL 6	0,984	0,973	0,812	0,612	0,529	0,948

Figure 19 shows the mean(|SHAP|) values. This method was used to provide an interpretation to the previously chosen models. As a matter of fact, thanks to these graphs we can see the importance that each variable assumes in the model.

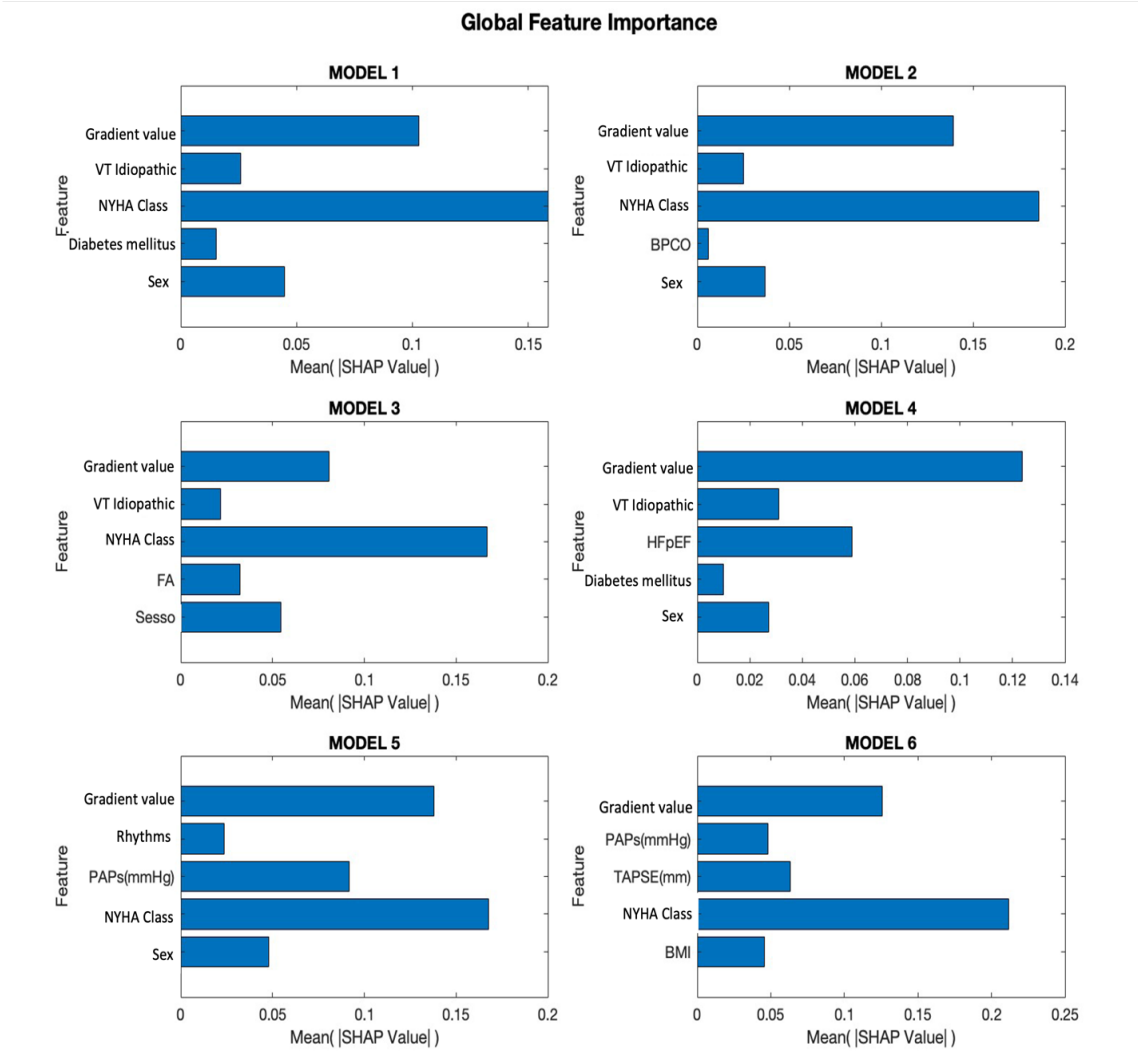

Figure 19. *| For each artificial neural network model, the SHAP values of the features are reported. Specifically, the mean of the absolute value.*

5. DISCUSSION

From Table 4 it can be seen that the most present variables are: NYHA Class, PAPs, Gradient value and Arrhythmic storm. In particular, the first three are present in all models, the last is present in four out of six models. This means that in the last four models, once the four variables listed have been fixed, the fifth variable is interchangeable between the various models, these are: HFmEF, HFrEF, Percentage LAT area, and Sex. In models two and six, however, the Arrhythmic storm variable (present in model six) is interchangeable with the Creatinine variable present in model two. From Table 6 we can obtain a comparison between the variables. The coefficients are all comparable to each other because multiplied by the assigned variable they have approximately the same orders of magnitude. This is important because sometimes the coefficients can be larger than others but then multiplied by the values of their variable, they turn out to have a lower impact. For example, in model two the coefficient of Creatinine is worth -1.133 and that of the PAPs variable is worth 0.279. In absolute value it seems that the variable Creatinine has a greater incidence than PAPs, but if we consider that in our dataset Creatinine has a range between [0.06 – 2.4], while PAPs have a range between [15 – 60], it is noticeable that the incidence of the PAPs variable is approximately five times greater than the Creatinine one. Among the four most recurring variables, the first three, i.e., those that are always present, also have a high incidence, while Arrhythmic storm, despite being very present, does not have an incidence as great as the others. The Sex variable, when present, has a greater impact than the others. The HFmEF variable in the model in which it appears has a high incidence compared to the other variables. Both the Sex variable and HFmEF are inversely proportional to output. In general, all the variables have the same orders of magnitude, so we cannot speak of a greater importance than the other variables, the most we can say is that in some models some variables matter a little more. Referring to Table 7 we can see that the variables PAPs, Arrhythmic storm and Gradient value are present in all six models. The TAPSE variable is present half of the time, that is, in three out of six models. In these cases, once these first four values have been fixed, we note that the variables FE, Rhythm and T-wave inversion are interchangeable with each other, respectively in the first three models. In the last two models, however, it is noted that the Percentage uni/bi potential and Percentage LAT area variables deriving from the analysis of the electroanatomical mapping system are interchangeable. As well as the TAPSE and TV Paroxysmal variables if the variable T-wave

inversion is fixed. As regards the coefficients, consider Table 9. First of all, it can be said that the PAPs variable is the one with the greatest impact compared to the other variables in each model. In the first model in particular, the order of magnitude reached by multiplying the range of values [15 – 60] by the assigned coefficient is greater than for the other variables. In the other models, however, it takes on approximately the same order of magnitude as the other variables, so we cannot speak of excessive relevance even though it is a little more influential than the other parameters. The Gradient value coefficients are always low compared to the others, as in the case of regressions, but having a range between [0 – 41] it is the second most influential variable after PAPs. Even in this case, anyhow, we are talking about an incidence that is a little greater than the others but not of greater relevance because the order of magnitude achieved is always approximately the same as that of the other variables. In the case of linear analyses, we have that six variables: PAPs, Rhythm, Percentage LAT area, Gradient value, Creatinine, and Arrhythmic storm, are present in both cases. For both regressions and linear SVM there are five non-recurring variables in both methods. As regards the variables in common between the two analyses, it can be said that they always have approximately the same coefficient values, i.e., between the various models if a range is established for the value of the coefficients of each individual variable, these are comparable between the two techniques. In addition to the same range, they also have the same sign, that is, if the coefficients of a variable are always positive in the regressions, they are also always positive in the linear SVM. This means that they always influence the output in the same direction and with the same weight. The results of the linear SVM therefore confirm those obtained from the regressions but also adding other variables. The most frequent variables, analyzing Table 10, are: Gradient value, Bipolar endocardial scar area, VT Idiopathic, and BPCO. The first three variables are present in all the selected models, while the last one is present in four out of six models. Once these four variables have been fixed for models one, three, four and six, the variables FA, BMI, Diabetes and Rhythm are interchangeable with each other. In the case of models one and two, if the FA variable is fixed, in addition to the three always present, the BPCO and LV aneurysm variables are interchangeable. Instead, by analyzing Figure 18 we can get an idea of how much these variables weigh on the output. Not having coefficients, it is not possible to have an equation that explains the model, but it is still possible to interpret it using the SHAP values. These allow us to get an idea of the global importance of the features. Gradient value has the greatest weight in all models. The other variables, excluding those with minimum

weight, fluctuate more or less in the same range of values. The Bipolar endocardial scar area variable is the one that seems to have the greatest relevance among the variables with an intermediate value range. VT Idiopathic, however, always has a fair incidence when it is present, only in model 2 does the FA variable appear to have a greater weight than VT Idiopathic. Among the variables with minimum weight are BPCO present in four models and with low relevance, the variables Family history of MCI and HFmEF also seem to have low relevance when they are present. The others have a minimal impact on the output, but still proportional to those with slightly higher values. It can be seen that the most present variables are: Gradient value, NYHA Class, Sex and VT Idiopathic. In particular, the Gradient value variable is present in every model, while Sex and NYHA Class occur in five of the six selected models, and VT Idiopathic in four of the six overall. In models one, two and three, if we fix the variables Sex, NYHA Class, VT Idiopathic and Gradient value we can see that the variables Diabetes, BPCO and FA are interchangeable with each other. If we also consider models one and four, we can see that the interchangeable variables are NYHA Class and FA. In the last two models the variables NYHA Class, Gradient value and PAPs recur while Sex and Rhythm are replaced by BMI and TAPSE. Consider Figure 19, in terms of relevance, the variable that prevails most is NYHA Class in the models in which it is present (1,2,3,5,6), with a value always close to 0.15. Another variable that is relevant is Gradient value which takes on the highest value in model 4, while it is only lower than NYHA Class in the others. The PAPs variable in model 5 has a significant impact, although still lower than NYHA Class and Gradient value. The Sex variable always takes constant values around 0.05. As regards the other variables present, we can note that they take on similar values to each other, hence with a similar incidence with respect to the output. We decided to implement different ML techniques so as to be able to find not only the combinations with the most correlated variables but also which of the different algorithms was most suitable for our database and which therefore gave us the best performance. We used different indices to evaluate the different performances obtained between the various techniques used. In particular, we calculated the AREA UNDER THE CURVE(AUC), Accuracy, Precision, Sensibility and Specificity. Each of these indices takes on a specific meaning that allows us to understand if the algorithm is working correctly and with good results. As we can see from Tables 5, 8, 11 and 13 in the first two columns we have the AUC values of testing and training respectively. Initially, we considered all values higher than 0.9 significant, however we can see that already with the linear regressions the maximum AUC

value for the test far exceeds 0.9 reaching up to 0.975. In particular, we achieved up to 113 combinations with AUC for the test above 0.9. With the linear Support Vector Machine, however, we can see that the maximum performance increases up to 0.992, and that we obtain around 3000 models with AUC for the test higher than 0.9, indicating a notable increase in significant combinations. As regards the non-linear Support Vector Machine and Artificial Neural Networks, the maximum AUC value we obtain for the test is 0.986, with thousands of models above 0.9. As regards the other indices, Accuracy measures the percentage of correct predictions compared to the total predictions made by the model and in our case takes values that oscillate between approximately 81% and 89% in each algorithm used. This indicates that all the different methodologies are able to correctly predict a good part of the database. Precision measures the percentage of predictions identified as positive that are actually positive. In our case we have values between 60% and 73%, indicators of good correctness in the prediction of positive instances. The Sensibility values are around 0.3. This indicates that the model is missing many of the true positive instances, producing a high number of false negatives. While the Specificity values are very good between 0.94 and 0.97. A high specificity value means that the model is effective in minimizing false positives, i.e., in not misclassifying negative instances as positive. This low sensitivity scenario can occur when the model as in our case does not have enough data to learn and therefore is unable to effectively discriminate between classes. In conclusion, we can say that already with linear regressions we obtain simple and easily interpretable models, with excellent performance. Performance which we significantly increase by using the linear Support Vector Machine, which is effective and at the same time capable of returning direct correlations between input and output variables. The last two techniques that use non-linear methods, thus, were tested even though we already had good results, also obtaining good performances which can certainly increase as the database increases. Therefore, when comparing performances, it is currently preferred to use linear methods since they are directly interpretable with the same performances. As a matter of fact, both linear regressions and linear Support Vector Machine give us coefficients that directly correlate the variables and the output. The non-linear Support Vector Machine and Artificial Neural Networks, anyhow, do not provide a direct explanation of the impact of each variable, which is why we used the SHAP values to attribute the correct relevance to each variable. The meaning of the most relevant features according to this study will be specified below. The NYHA Class, acronym for New York Heart Association, is a classification used in medicine to evaluate

the severity of symptoms and limitations of heart failure. Heart failure is a condition in which the heart is unable to pump enough blood to meet the body's needs. The NYHA classification divides patients into four classes based on severity of symptoms and limitation of physical activity. Pulmonary arterial pressure (PAP) is the measurement of blood pressure in the pulmonary arteries, the blood vessels that carry blood from the heart to the lungs for oxygenation. It is important in evaluating heart and lung function, although in itself it is not directly related to cardiac arrhythmias. The term " arrhythmic storm " refers to a condition in which a person experiences a series of cardiac arrhythmias in rapid succession or persistently. In other words, it is a period in which frequent episodes of heart rhythm disorders occur, even serious ones. TAPSE is an acronym that stands for " Tricuspid Cancel Plane Systolic Excursion ". This is a parameter used to evaluate the function of the right ventricle of the heart. The measurement of TAPSE is obtained using echocardiography, during an echocardiographic examination, the images of the movement of the tricuspid annulus in the right ventricle during contraction are acquired. Hence, it provides information on the contraction and movement of the tricuspid annulus, helping to evaluate the function of the right ventricle of the heart. The acronym "idiopathic VT" refers to "idiopathic ventricular tachycardia". A type of cardiac arrhythmia characterized by abnormally fast heartbeats originating from the ventricles of the heart. The designation "idiopathic" indicates that the specific cause of ventricular tachycardia is not known or is not clearly identifiable. The acronym "BPCO" (BPCO) stands for Chronic Obstructive Pulmonary Disease, a chronic lung disease characterized by progressive obstruction of the airways. It is not directly related to arrhythmology, but it can affect cardiac function, for example, as pulmonary problems can alter pulmonary arterial pressure and blood oxygenation. The acronym Bipolar endocardial scar area refers to the “endocardial scar area,” an area of scar or scar tissue that develops on the inner surface of the heart. Endocardial scarring can form following several events, such as a myocardial infarction or cardiac surgery. When heart tissue is damaged, the healing process can lead to the formation of scar tissue, which is stiffer and less conductive than normal heart tissue. This can affect the electrical conductivity of the heart and contribute to the development of arrhythmias. In particular, the endocardial scar area may be involved in the formation of reentrant circuits that are associated with certain arrhythmias, such as ventricular tachycardia. The scar can create a pathway through which electrical impulses can circulate abnormally, causing irregular heart rhythms. Evaluation of the endocardial scar area can be performed through various diagnostic techniques, such as

echocardiography, cardiac magnetic resonance imaging or computed tomography. As regards the variables extracted from the mapping system, we want to briefly discuss the results. A novelty of this project is the insertion of parameters extracted from a mapping analysis electroanatomical, the values taken into consideration are not directly analyzed by the machine, but have been calculated with a specific ad hoc algorithm. First of all, we want to remember the meaning of these values. The variable Percentage LAT area is expressed as a percentage and refers to the extension of the delay area, the variable Gradient value expresses in milliseconds the value of the difference between the zone with the greatest delay and the zone with the greatest advance when these are sufficiently close, and the variable Percentage uni/bi potential, also expressed as a percentage, refers to the extent of the area composed of points where the difference between the bipolar and unipolar potential is significantly high. Gradient value is present in all the analyzes carried out, and specifically in all the best six models for each analysis. The other two are present only in linear relationships and disappear in non-linear ones, in particular Percentage LAT area is present both in Regressions and in linear SVM, Percentage uni/bi potential only in linear SVM. However, we would like to remember that in this study the first six models were taken into consideration for each analysis; from the processing carried out, many others emerge with similar or slightly lower performances compared to those chosen, in which the variables Percentage LAT area and Percentage uni/bi potential are present with greater frequency. We cannot therefore exclude the importance of these two variables, even if they were not very present in this study. Nonetheless, one cannot fail to notice the great relevance of the parameter referring to the gradient.

6. CONCLUSIONS

The present thesis aimed to provide engineering support in the search for predictive variables for relapse. The goal was to analyze which features could have a greater influence than others on prediction, compare the performance of various models obtained, and repeat the entire process for multiple machine learning methods to determine the best-suited one in this case. A final, even though not less important, objective was to understand the impact of certain parameters derived from the electroanatomic mapping system. The models' performances in all the machine learning techniques used are good, as all variables included in the study were classified as important for this purpose. In this thesis, 220 patients were analyzed, resulting

in a low number of observations. For this reason, it is not possible to speak of scientific evidence but rather of a promising prototype that, from preliminary analyses, already shows good performance values. The AUC and other parameters are good; the only one with low values is Sensibility, precisely due to the low number of observations analyzed. Among all machine learning techniques, the linear support vector machine is preferred because the results are promising, and it remains a method with greater interpretability even with similar performance to other techniques. The variables identified as relevant are parameters that can be monitored but are difficult or sometimes impossible to intervene upon. It is therefore not possible to avoid or prevent a relapse, but rather to predict it. The novelty of this project lies in combining these classic analyses, already widely researched, with parameters from the mapping system. It is neither possible nor meaningful to include point-to-point values obtained from the analysis and identifiable in the export; instead, the focus is on analyzing the values of combinations of these parameters. In this case, parameters related to the bipolar and unipolar potential difference and two parameters related to delay zones were calculated. Specifically, one parameter was related to the extension of the zone, and the other to the difference in delay and advance of two adjacent zones. These parameters were extracted and calculated with a specially designed algorithm, and many others can still be calculated and added. Even in this preliminary analysis, these values have proven to be useful and influential for relapse, sometimes with high relevance. We are confident that additional parameters will be found to add to the relapse prediction analyses, and many of these will be closely related to it, as already demonstrated in this work. The analyses are preliminary, but with an increase in the number of patients and thus the number of observations, more interesting results will be possible. Nevertheless, it remains a promising prototype for future, with more truthful and accurate results.

7. BIBLIOGRAPHY

[1] Mitchell, L. Brent. “Panoramica Sulle Aritmie - Disturbi Dell’apparato Cardiovascolare.” Manuali MSD Edizione Professionisti, Manuali MSD, 20 Sept. 2023, www.msdmanuals.com/it-it/professionale/disturbi-dell-apparato-cardiovascolare/panoramica-sulle-aritmie-e-i-disturbi-della-conduzione/panoramica-sulle-aritmie.

[2] Huizar JF, Ellenbogen KA, Tan AY, Kaszala K. Arrhythmia-Induced Cardiomyopathy: JACC State-of-the-Art Review. J Am Coll Cardiol. 2019 May 14;73(18):2328-2344. doi: 10.1016/j.jacc.2019.02.045. PMID: 31072578; PMCID: PMC6538508.

[3] Sossalla S, Vollmann D. Arrhythmia-Induced Cardiomyopathy. Dtsch Arztebl Int. 2018 May 11;115(19):335-341. doi: 10.3238/arztebl.2018.0335. PMID: 29875055; PMCID: PMC5997886.

[4] Gopinathannair R, Etheridge SP, Marchlinski FE, Spinale FG, Lakkireddy D, Olshansky B. Arrhythmia-Induced Cardiomyopathies: Mechanisms, Recognition, and Management. J Am Coll Cardiol. 2015 Oct 13;66(15):1714-28. doi: 10.1016/j.jacc.2015.08.038. PMID: 26449143; PMCID: PMC4733572.

[5] De Ferrari, G. M. Cos’ è lo scompenso cardiaco I sintomi e le cause cardiache. PREVENZIONE COME CURA, 4.

[6] McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M, Burri H, Butler J, Čelutkienė J, Chioncel O, Cleland JGF, Coats AJS, Crespo-Leiro MG, Farmakis D, Gilard M, Heymans S, Hoes AW, Jaarsma T, Jankowska EA, Lainscak M, Lam CSP, Lyon AR, McMurray JJV, Mebazaa A, Mindham R, Muneretto C, Francesco Piepoli M, Price S, Rosano GMC, Ruschitzka F, Kathrine Skibelund A; ESC Scientific Document Group. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J. 2021 Sep 21;42(36):3599-3726. doi: 10.1093/eurheartj/ehab368. Erratum in: Eur Heart J. 2021 Oct 14;: PMID: 34447992.

[7] Heidenreich, P. A., Bozkurt, B., Aguilar, D., Allen, L. A., Byun, J. J., Colvin, M. M., Deswal, A., Drazner, M. H., Dunlay, S. M., Evers, L. R., Fang, J. C., Fedson, S. E., Fonarow, G. C., Hayek, S. S., Hernandez, A. F., Khazanie, P., Kittleson, M. M., Lee, C. S., Link, M. S., … Yancy, C. W. (2022). 2022 AHA/ACC/HFSA guideline for the management of heart failure: Executive summary: A report of the american college of cardiology/american heart association joint committee on clinical practice guidelines. Circulation, 145(18). https://doi.org/10.1161/cir.0000000000001062

[8] Istat, “Mortalità per Territorio Di Evento.” Mortalità per Territorio Di Evento, 2020, dati.istat.it/Index.aspx?DataSetCode=DCIS_CMORTE1_EV.

[9] Malattie Cardiovascolari - Salute.Gov.It, www.salute.gov.it/imgs/C_17_navigazioneSecondariaRelazione_1_listaCapitoli_capitoliItemName_1_scarica.pdf.

[10] Gopinathannair, R., Etheridge, S. P., Marchlinski, F. E., Spinale, F. G., Lakkireddy, D., & Olshansky, B. (2015). Arrhythmia-induced cardiomyopathies: mechanisms, recognition, and management. Journal of the American College of Cardiology, 66(15), 1714-1728.

[11] MEDI, Caroline, et al. Tachycardia-mediated cardiomyopathy secondary to focal atrial tachycardia: long-term outcome after catheter ablation. Journal of the American College of Cardiology, 2009, 53.19: 1791-1797.

[12] Elming, Marie Bayer, Sophia Hammer-Hansen, Inga Voges, Evangelia Nyktari, Anna Axelsson Raja, Jesper Hastrup Svendsen, Steen Pehrson et al. "Right ventricular dysfunction and the effect of defibrillator implantation in patients with nonischemic systolic heart failure." Circulation: Arrhythmia and Electrophysiology 12, no. 3 (2019): e007022.

[13] Amin, A.K., Gold, M.R., Burke, M.C., Knight, B.P., Rajjoub, M.R., Duffy, E., Husby, M., Stahl, W.K. and Weiss, R., 2019. Factors associated with high-voltage impedance and subcutaneous implantable defibrillator ventricular fibrillation conversion success. Circulation: Arrhythmia and Electrophysiology, 12(4), p.e006665.

[14] Daniele Muser et al. «Role of cardiac imaging in patients undergoing catheter ablation of ventricular tachycardia». In: Journal of Cardiovascular Medicine 22.10 (ott. 2021), p. 727. issn: 1558-2027. doi: 10.2459/JCM.0000000000001121. url: https://journals.lww. com / jcardiovascularmedicine / pages / articleviewer . aspx ? year = 2021 & issue = 10000&article=00001&type=Fulltext.

[15] Samuel J. Asirvatham e Matthew J. Swale. «Imaging and Cardiac Ablation». In: JACC: Cardiovascular Imaging 4.7 (lug. 2011), pp. 727–729. issn: 1936878X. doi: 10.1016/ j.jcmg.2010.12.010. url: https://linkinghub.elsevier.com/retrieve/pii/ S1936878X11003536.

[16] Calkins H, Sousa J, El-Atassi R, et al. Diagnosis and cure of the Wolff– Parkinson–White syndrome or paroxysmal supraventricular tachycardias during a single electrophysiologic test. N Engl J Med 1991;324:1612-8.

[17] Haines DE, Watson DD, Verow AF. Electrode radius predicts lesion ra- dius during radiofrequency energy heating: validation of a proposed ther- modynamic model. Circ Res 1990;67:124-9.

[18] Simmers TA, Wittkampf FHM, Hauer RNW, Robles de Medina EO. In vivo ventricular lesion growth in radiofrequency catheter ablation. Pac- ing Clin Electrophysiol 1994;17:523-31.

[19] Langberg JJ, Calkins H, el-Atassi R, et al. Temperature monitoring dur- ing radiofrequency catheter ablation of accessory pathways. Circulation 1992;86:1469-74.

[20] Calkins H, Prystowsky E, Carlson M, Klein LS, Saul JP, Gillette P. Temperature monitoring during radiofrequency catheter ablation proce- dures using closed loop control. Circulation 1994;90:1279-86.

[21] Reddy VY, Reynolds MR, Neuzil P, et al. Prophylactic catheter ablation for the prevention of defibrillator therapy. N Engl J Med 2007; 357: 2657-65.

[22] Andrea Saglietto, Fiorenzo Gaita, Carina Blomstrom-Lundqvist, Elena Arbelo, Nikolaos Dagres, Josep Brugada, Aldo Pietro Maggioni, Luigi Tavazzi, Josef Kautzner, Gaetano Maria De Ferrari, Matteo Anselmino, on behalf of the AFA LT registry investigators group, AFA-Recur: an ESC EORP AFA-LT registry machine-learning web calculator predicting atrial fibrillation recurrence after ablation, EP Europace, Volume 25, Issue 1, January 2023, Pages 92–100, https://doi.org/10.1093/europace/euac145

[23] Cronin EM, Bogun FM, Maury P, Peichl P, Chen M, Namboodiri N, Aguinaga L, Leite LR, Al-Khatib SM, Anter E, Berruezo A, Callans DJ, Chung MK, Cuculich P, d'Avila A, Deal BJ, Bella PD, Deneke T, Dickfeld TM, Hadid C, Haqqani HM, Kay GN, Latchamsetty R, Marchlinski F, Miller JM, Nogami A, Patel AR, Pathak RK, Saenz Morales LC, Santangeli P, Sapp JL Jr, Sarkozy A, Soejima K, Stevenson WG, Tedrow UB, Tzou WS, Varma N, Zeppenfeld K. 2019

HRS/EHRA/APHRS/LAHRS expert consensus statement on catheter ablation of ventricular arrhythmias: Executive summary. J Arrhythm. 2020 Jan 3;36(1):1-58. doi: 10.1002/joa3.12264. PMID: 32071620; PMCID: PMC7011820.

[24] Ziad Issa JMM, Douglas P. Zipes. Clinical Arrhythmology and Electrophysiology: A Companion to Braunwald's Heart Disease. 1 ed. Philadelphia, PA: Saunders, an imprint of Elsevier Inc.; 2009

[25] Smeets JL, Ben-Haim SA, Rodriguez LM, Timmermans C, Wellens HJ. New method for nonfluoroscopic endocardial mapping in humans: accuracy assessment and first clinical results. Circulation. 1998 Jun 23;97(24):2426-32. doi: 10.1161/01.cir.97.24.2426. PMID: 9641695.

[26] Bunch TJ, Weiss JP, Crandall BG, et al. Image integration using intracardiac ultrasound and 3D reconstruction for scar mapping and ablation of ventricular tachycardia. Journal of cardiovascular electrophysiology 2010;21:678-84.

[27] Webster B. Carto 3 System Fact Sheet. https://wwwbiosensewebstercom/documents/carto3-factsheetpdf?Cache=1%2F19%2F2015+3%3A56%3A28+PM 2014.

[28] Arlot, S. and Celisse, A., 2010. A survey of cross-validation procedures for model selection.

[29] Altman, D.G. and Bland, J.M., 1994. Diagnostic tests. 1: Sensitivity and specificity. BMJ: British Medical Journal, 308(6943), p.1552

[30] Sadegh-Zadeh, Kazem. "In dubio pro aegro." Artificial Intelligence in Medicine 2, no. 1 (1990): 1-3.

[31] Turing, Alan M. Computing machinery and intelligence. Springer Netherlands, 2009.

[32] Breiman, L., 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science, 16(3), pp.199-231.

[33] James G, Witten D, Hastie T, Tibshirani R, An Introduction to Statistical Learning: With Applications in R. New York: Springer; 2013. [Google Scholar]

[34] Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York, NY: Springer; 2009. [Google Scholar]