



UNIVERSITÀ POLITECNICA DELLE MARCHE
FACOLTÀ DI INGEGNERIA

Master's Degree in BIOMEDICAL ENGINEERING

**EVALUATION OF UNSUPERVISED MACHINE LEARNING AS
A SUPPORT IN PRE-HOSPITAL ELECTROCARDIOGRAPHY**

Supervisor: Dr.ssa Agnese Sbröllini

Co-supervisors: Prof.ssa Laura Burattini

Prof. Cees A. Swenne

Candidate: Maria Sampiero

Academic Year 2023/2024

Abstract

Cardiovascular diseases (CVDs) remain the leading cause of disease burden globally, contributing to premature mortality, disability, and rising health care costs. In this spectrum of cardiac conditions, myocardial ischemia and acute coronary syndrome (ACS) are particularly notable, which are joined by almost the same causes, with the atherosclerosis being the most prevalent. The latter has as main consequence the intraluminal thrombi formation, which occludes the blood vessel resulting in acute myocardial ischemia. When occurring in the coronary circulation, it manifests as ACS. Prompt decisions and accurate diagnosis of these conditions, already in the pre-hospital phase, are crucial for preserving cardiac function as much as possible. In the emergency department (ED), electrocardiography (ECG) is the most widely used initial diagnostic test to support clinical diagnosis and aid in risk stratification, as recommended by the current clinical guidelines for screening patients presented with chest pain and anginal equivalents. The standard 12-lead ECG is the criterion standard for electrocardiographic detection of acute myocardial ischemia/injury and is reported to be the single most important method to rapidly identify ACS in the ED. Nevertheless, due to the unpredictable and dynamic nature of acute ischemic changes, a single ECG snapshot may be inadequate. Thus, it is recommended a comparison of the acute ECG, which is under suspicion, to a previously recorded non-acute ECG of the same patient, to account for interindividual variability. With the improvement in pattern recognition abilities of artificial intelligence and the urgency of large-scale data management, the unsupervised machine learning (ML) can be employed for quickly discovering data distributions and relevant trends, learning the underlying structures, facilitating diagnosis-making procedures, without any predefined labels. The analysis focuses on ten clustering methods, which are applied to data from SUBTRACT study, which are characterized by class imbalance. More specifically, the dual purpose of this thesis consists of: (1) demonstrating the effectiveness of the clustering techniques in revealing the underlying structure of data by correctly distinguishing pathological cases from healthy individuals; (2) evaluating the role of both ECG features set, 18 direct measurements and 28 serial features, on the performance of those algorithms and their ability to support accurate diagnosis of myocardial ischemia and ACS. First, the comparative analysis of clustering methods, assessed through evaluation metrics, reveals that the implemented techniques perform well with both databases, especially with myocardial ischemia database, achieving high accuracy scores exceeding 70,33%. Notably, the CLARA methodology results in the most effective clustering method across both conditions and feature sets; with accuracy and F1 score consistently above

74,65% and 69,75%, respectively. Secondly, in ACS database the serial features are found to be less relevant for the diseases detection, differently from the myocardial ischemia, where some algorithms benefit from their inclusion. These findings underscore the importance of database characteristics, within parameters choice, in influencing the algorithm performance. In summary, integrating advanced ML methods into routine diagnostic processes for ECG recordings in pre-hospital care has the potential to support healthcare professionals in quickly analyzing medical reports and making diagnoses, thereby reducing the risk of health complications and ultimately improving patient outcomes.

INDEX

Introduction

Chapter 1 – Myocardial Ischemia & Acute Coronary Syndrome

- 1.1. Epidemiology
- 1.2. Heart Anatomy and Physiological Responses of the Coronary Circulation
- 1.3. Electrocardiography
- 1.4. Acute Myocardial Ischemia: pathophysiology, clinical presentation, and underlying mechanisms
- 1.5. Acute Coronary Syndrome: pathophysiology, clinical presentation, and underlying mechanisms

Chapter 2 – Triage & Assessment of syndromes

- 2.1 Triage
- 2.2 Electrocardiography as diagnostic tool
- 2.3 Serial electrocardiography

Chapter 3 – Unsupervised Machine Learning

- 3.1. Machine Learning and general workflow
- 3.2. Unsupervised Machine Learning
- 3.3. Clustering
 - 3.3.1. K-means
 - 3.3.2. K-medoids
 - 3.3.3. Spectral Clustering
 - 3.3.4. Agglomerative Clustering
 - 3.3.5. CLARA
 - 3.3.6. Fuzzy C-means
 - 3.3.7. Gaussian Mixture Model
 - 3.3.8. BIRCH
 - 3.3.9. Self-organizing Map
 - 3.3.10. Mean Shift

Chapter 4 – Materials & Method

- 4.1 Database
- 4.2 Implementation
- 4.3 Statistical Analysis

Chapter 5 – Results

Chapter 6 – Discussion

Chapter 7 – Conclusion

References

Introduction

The World Health Organization (WHO) identifies cardiovascular diseases (CVDs) as the primary cause of death globally. An estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths [1]. Cardiometabolic, behavioral, environmental, and social risk factors are major drivers of CVD [2]. Atherosclerosis is the most prevalent form of CVD, defined by the formation of cholesterol-rich plaques within the inner arterial walls, which results in turns in intraluminal thrombus development. This condition may lead to a partial or complete occlusion of the blood vessel, causing acute myocardial ischemia and the subsequent cell death in the areas supplied by itself. When occurring in the coronary circulation, this scenario called acute coronary syndrome (ACS). If the blockage of the blood flow persists without spontaneous resolution or timely effective treatment, a myocardial infarction can develop within hours [1]. Therefore, prompt decisions and correct diagnosis of these diseases, already in the pre-hospital phase, are of pivotal importance for preventing or reducing functional and structural myocardial damages.

Upon the arrival of the medical emergency services, an electrocardiogram (ECG) is routinely made, and its interpretation plays an essential role in the triage decisions about the patient's actual health status and subsequent required treatment. Although 12-lead ECG considered the key objective prehospital diagnostic tool for the initial evaluation of patients with suspected symptoms of myocardial ischemia and ACS, the complex nature of chest pain combined with unpredictable and dynamic acute ischemic changes suggests that a single ECG snapshot is inadequate. Thus, it was introduced the serial electrocardiography, which combines the acute ECG and a previously obtained (reference) non-acute ECG of the same patient, allowing to identify ischemia-induced electrocardiographic changes by correcting for interindividual ECG variability [1].

With the improvement in pattern recognition abilities of artificial intelligence, it is necessary to reconsider how the reference ECG and the contextual patient information when presenting chest pain should be combined to get the best possible initial working diagnosis. More specifically, the focus of this work is on the potential integration of unsupervised machine learning (ML) as a promising approach to enhance and support diagnostic accuracy in emergency medical settings, as the increasing demand for rapid and accurate cardiac assessment. Thus, the dual purpose of this work consists of: (1) demonstrating the effectiveness of the clustering techniques in revealing the underlying structure of data by correctly distinguishing pathological cases from healthy individuals; (2) evaluating the role of both ECG feature classes, 18 direct measurements and 28

serial features, on the performance of those algorithms and their ability to support accurate diagnosis of myocardial ischemia and ACS conditions.

This thesis organized in the following manner: *Chapter 1* focuses on the epidemiological and physiological background needed to contextualize the problematics, while *Chapter 2* on the triage procedures in the emergency department (ED) for the assessment and the management of patients with symptoms suspected to be related to both syndromes. Passing to the more technical concepts, *Chapter 3* introduces machine learning (ML) concept, outlines a general workflow, a description of unsupervised ML and then moves on to discuss clustering algorithms relevant to this analysis. *Chapter 4* details the characteristics of the database used, the implementation process in Python within a Google Colab environment, and the statistical analysis conducted. Finally, the work concludes with the presentation of the results in *Chapter 5*, a discussion of their implications in *Chapter 6*, and the overall conclusions in the final *Chapter 7*.

Chapter 1 – Myocardial Ischemia & Acute Coronary Syndrome

1.1. Epidemiology

According to the World Health Organization, cardiovascular diseases (CVDs) are the leading cause of death globally [1]. The estimates by the Global Burden of Disease (GBD) Study 2019 [2], an ongoing multinational collaboration, aimed at providing consistent population health trend over time, highlighted that: prevalent cases of total CVD nearly doubled from 271 million in 1990 to 523 million in 2019, and the number of CVD deaths steadily increased from 12.1 million to 18.6 million. CVD was the cause of 6.2 million deaths occurring between the ages of 30 and 70 years in 2019. Prevalent cases of total CVD are intended to rise as the result of population growth and aging, especially in Northern Africa, Asia, Latin America, and the Caribbean, where the share of older persons is estimated to double between 2019 and 2050. Alarmingly, CVD burden continues its decades-long rise for almost all countries with the exception of high-income countries, and the age-standardized rate of CVD has begun to increase especially in low- and middle-income countries, as shown on *Fig. 1*, reflecting the disparity in terms of access to healthcare, including effective primary and secondary prevention strategies, as well as CVD risk factors exposure [2].

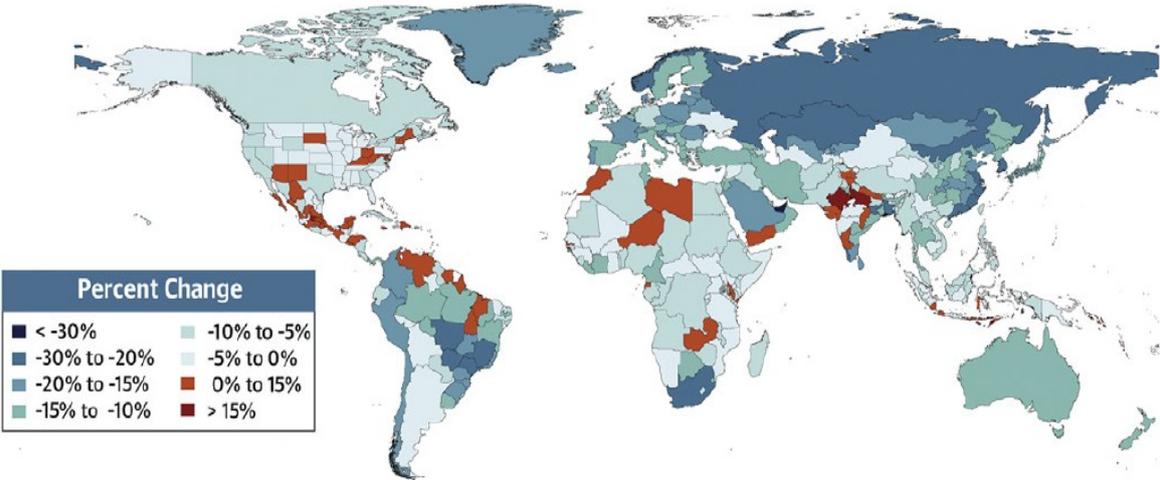


Fig. 1 – Percent Change in Age-Standardized CVD Death Rate from 2010-2019. From [2].

Global patterns of total CVD have significant implications for clinical practice and public health policy development. In fact, in the USA between 2010 and 2030 the total direct medical costs for CVD are projected to triple from \$273 billion to \$818 billion [3].

Ischemic heart disease (IHD) is the most common cause of CVD death, accounting for 38% of all CVD deaths in females and 44% in males, proving to be a major threat to public health [4]. The increasing incidence of IHD expected to continue, due not only to population aging, but also to the increased prevalence of obesity, diabetes, and metabolic syndrome [5].

Acute coronary syndrome (ACS) is often the first clinical manifestation of CVD, causing in the USA, more than one million cases every year [3]. Despite the improvement in survival associated with ACS, this medical condition continues to be associated with fatal outcomes and place a burden on the entire health care system. Of 1.2 million patients suffering from ACS or cardiac death each year, more than half die either before reaching the hospital or in the Emergency Department (ED) [6].

Health systems and countries need to focus on delivering effective interventions that will reverse these trends, by improving diet and physical activity, pre- and in-hospital care, survival and quality of life, since a prompt diagnosis and an accurate therapeutic approach are of paramount importance to save the long-term prognosis of the heart [2].

1.2. Heart Anatomy and Physiological Responses of the Coronary Circulation

The heart is composed of muscle tissue known as myocardium, which rhythmically contracts to circulate blood throughout the body. It consists of four cavitory chambers, whose walls consist of a mechanical syncytium of myocardial cells, called cardiomyocytes. At the exit of each chamber a valve closes after contraction to prevent significant retrograde flow when the chamber relaxes, thereby ensuring that downstream pressures do not exceed chamber pressures. The right heart includes the atrium, which receives blood from most of the body, leading into a larger right ventricle through the tricuspid valve. The right ventricle then pumps blood through the pulmonary valve into the lungs, where the blood oxygenated and relieved of carbon dioxide. The left atrium and left ventricle form the left heart. The left atrium receives blood from the lungs and, passing

through the bicuspid valve, directs it into the left ventricle, which propels the blood through the aorta to the rest of the body.

The heart functions as a complex electrochemical machine, where the generation of electrical impulses is essential for triggering the contractile process. Each mechanical heartbeat, or systole, is triggered by an action potential which originates from a rhythmic pacemaker located in the sinoatrial node of the heart and is conducted rapidly throughout the organ to produce a coordinated contraction. Thus, before the action potential can propagate, it must be initiated by pacemakers, cardiac cells that possess the property of automaticity, the ability to spontaneously depolarize. As a result, a cardiac cycle is generated, effectively pumping blood throughout the body, defined as a sequence of alternating contraction and relaxation of the atria and ventricles. Each cardiac cycle consists of a diastolic phase (also called diastole), during which the heart chamber fills with blood that receives from the veins in a relaxed state, and a systolic phase (also called systole), when the heart chambers contract to pump blood towards the periphery via the arteries. Both atria and ventricles experience alternating states of systole and diastole, with the atria contracting before the ventricles begin their contraction under normal conditions [7]. Action potential is a complex event that involves movement of various ions across the cell membrane, either actively or passively. The changes of membrane potential modified by changing the extracellular and intracellular concentration of ions such as sodium (Na^+), calcium (Ca^{2+}), and potassium (K^+). Until the arrival of the electrical stimulus, the cell membrane maintains a stable negative potential at resting state, called resting membrane potential. When the membrane potential is elevated above a threshold potential, an abrupt increase in the membrane potential will occur ("depolarization") and be followed by a plateau of positive potential before the membrane potential gradually returns to the resting level ("repolarization"). Depolarization refers to the reduction in the degree of electronegativity in a resting cell, while repolarization refers to the return to resting potential. The cardiac action potential consists of five phases, as illustrated in *Fig. 2*:

- Phase 0 (depolarization): upon stimulation, the upstroke occurs via influx of Na^+ and the cell becomes positively charged.
- Phase 1 (early repolarization): K^+ channels open and a brief efflux repolarized the cell slightly.
- Phase 2 (plateau phase): almost simultaneous with the opening of K^+ channels in phase 1, persistent Ca^{2+} channels open. The steady influx of calcium into the cell gives a long duration to that phase, explaining the reason the vast majority of the ventricular myocardium contracts simultaneously.

- Phase 3 (repolarization): Ca^{2+} channels close and K^+ channels open again, allowing cell repolarization.
- Phase 4 (resting membrane potential): efflux of K^+ establishes a negative resting membrane potential, approximately -80/-90 mV.

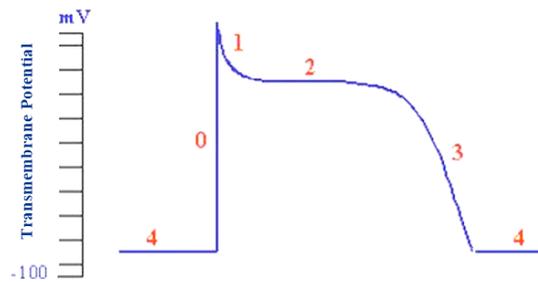


Fig. 2 – The action potential in contractile myocardial cells.

The myocardium is an aerobic tissue that requires continuous perfusion of oxygenated blood to generate the energy necessary for its contraction. In normal conditions, the heart spends much more oxygen (10 mL O_2 /min/100 at rest) compared to other organs, because of its continuous work [8]. Heart muscle exploits adenosine triphosphate (ATP), a high-energy phosphate molecule, as its primary fuel to guarantee the contraction-relaxation cycle. For supporting the heart functional demand, it is requested to human cardiac cells to produce massive quantities of ATP, thanks to an abundance of mitochondria, which comprise about 40–50% of the cardiac cellular mass [9]. Approximately 75% of myocardial energy used to sustain the contractile mechanism of the myocardium, mostly during the isovolumetric contraction and ejection phases of the cardiac cycle [10].

The performance of the heart as a pump is dependent primarily upon the contraction and relaxation properties of the myocardium. However, there are other factors, such as: the geometric organization of the myocardial cells, the properties of the cardiac tissue, the heart's electrical rhythm, valvular function, and the adequacy of delivery of oxygenated blood via the coronary arteries to meet the metabolic demands of the myocardium [7]. Different studies, summarized in [8], have demonstrated that the heterogeneous heart perfusion is closely related to the structural heterogeneity of terminal vasculature, implying that the metabolism heterogeneity may also depend on the tissue's adaptation to oxygen availability.

Under physiological circumstances, the coronary circulation matches blood flow with myocardial oxygen demand by coordinating the vascular resistances within microvasculature, where the endothelium plays a key role. Under baseline conditions, oxygen extraction from the arterial blood closed to 60% to 70%, and so an increase in myocardial oxygen demand can only be met by an adequate rise in coronary blood flow (CBF) [11]. To maintain an adequate oxygen and energetic substrate amount to every cardiomyocyte, coronary vessels show an advanced vascular network with extremely organized flow regulatory mechanisms. The cardiovascular system consists of a closed circuit of blood vessels. A continuous motion of blood from the left to the right heart provides the individual cells with sufficient nutrients. Blood flow in the vessels regulated by the myogenic and neurogenic processes: the first refers to the contraction and relaxation of smooth muscle in the vessel wall, whereas the latter is driven by the autonomic nervous system [7].

The coronary arterial system consists of three compartments, each with a distinct function, whose anatomic borders cannot be clearly defined in vivo.

- The *proximal* compartment represents by the epicardial arteries, characterized by multiple early intramural branches with a diameter ranged from 5.0 to 0.5 mm, which have a capacitance function and offer little resistance to CBF. During the systole, this first pattern accumulates elastic energy as they increase their blood content up to $\approx 25\%$. At the onset of diastoles, this elastic energy converted into kinetic one, that allows the prompt reopening of the intramyocardial vessels, compressed during systole. This mechanism has a particular influence, considering that 90% of CBF takes place in diastole.
- The *intermediate* compartment includes pre-arteriolar vessels, with diameters ranging from ≈ 500 to $100 \mu\text{m}$ and characterized by a measurable pressure drop along their length. It directed to the subendocardial myocardium and papillary muscles with less initial branching. Their primary function is to maintain pressure at the origin of downstream arterioles within a narrow range when changes in coronary perfusion pressure or flow take place.
- The *distal* compartment consists of the arterioles, characterized by short branches with diameters $<100 \mu\text{m}$ that supply mainly the subepicardial myocardium. Arterioles are the site where metabolic regulation of CBF takes place, since their tone influenced by substances produced by the surrounding cardiomyocytes [11].

Along the coronary tree, the distal vessels, arterioles, and capillaries constitute approximately 40–50% of total coronary resistance. Small arteries offer a small fraction of resistance, around 15–20%, and are more responsive to flow-dependent dilatation. Instead, large epicardial arteries

provide just a tiny fraction of resistance to coronary blood flow, as they are capacitance vessels, but more sensitive to changes in intravascular pressure, too [8].

Abnormal dilatory responses of the coronary micro-vessels, coronary microvascular spasm, and endothelium dysfunction have been identified as pathogenic mechanisms for CVDs, causing the impairment in the cross talk between myocardial energy state and CBF. There are major metabolic consequences caused by disruption or reduction of blood flow in coronary circulation and the resultant decrease in oxygen supply to the affected portion of myocardium (i.e. ischemia) and/or other types of lack of oxygen (i.e. hypoxia) in myocardial tissues caused by environmental factors or demand/supply imbalance [9].

1.3. Electrocardiography

Bio-signals are electrical, mechanical, thermal, measured over time from the human body. They became suitable for making medical diagnoses in 1895, when Willem Einthoven invented electrocardiography (ECG, or sometimes EKG) as a clinical usable, quasi-periodic and non-invasive device. An ECG device measures the electrical activity of the heart muscle and depicts the complete cardiac cycle on an individual heartbeat exploiting electrical polarization-depolarization patterns of the heart. ECG signals are comprised of the superposition of different action potentials from the heart beating. Indeed, the pattern of electrical propagation is not random, but spreads over the heart structure in an organized and coordinated manner, which leads to an effective systole [12]. An ECG denotes the body surface potentials, derived from current flows within the body, as a function of time. The simplest mathematical model for relating the cardiac generator to the body surface potential is the single dipole model. As an action potential propagates through a cell, there is an associated intracellular current generated in the direction of propagation, at the interface of resting and depolarizing tissue. So that, the elementary electrical source of the surface ECG is the current dipole [7].

Typical technical parameters of ECG-related signal are as follows: signal frequency ranges from 0.05 to 150 Hz, recording frequency ranges from 250 to 1000 Hz, amplitude levels range from 0.5 to 5 mV, and recording duration can vary from 10 seconds to 24 hours [12].

The ECG signal consists of P wave, QRS complex, T wave (and sometimes U, although this wave is often challenging to identify, as it may be absent, have a very low amplitude, or be masked by the next beat), as illustrated in *Fig. 3*. *Fig. 4* shows the RR interval, a measure of inter beat timing,

which is useful in the cardiac function assessment and in case of arrhythmia detection. The ECG recordings do not appear the same in all the leads used for the examination, since the ECG polarity and the shape of ECG waveforms may change depending on lead that considered [13].

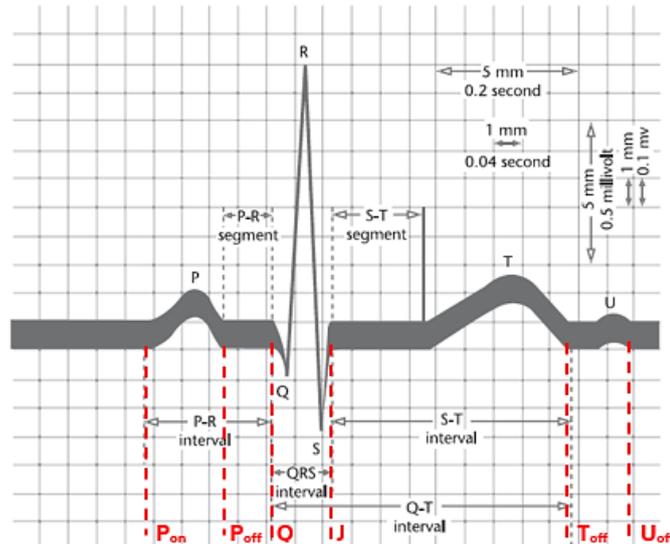


Fig. 3 – Normal features of the electrocardiogram. Modified from [7].

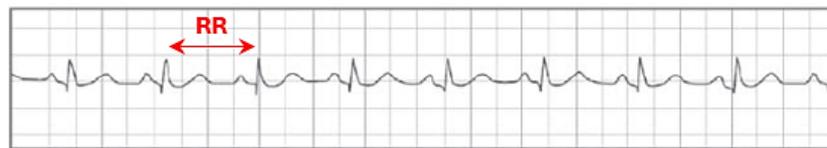


Fig. 4 – Normal sinus rhythm. Modified from [7].

These waves correspond to specific electric phenomena on the cardiac surface. Looking at Fig. 3, the cardiac cycle begins with the P wave (the relative start and end points indicated as P_{on} and P_{off}), which corresponds to the period of atrial depolarization in the heart. This followed by the QRS complex, which is the most recognizable feature of an ECG waveform, corresponding to the period of ventricular depolarization. The start and end points of the QRS complex referred to as the Q and J points. The T wave follows the QRS complex and coincides with the period of ventricular repolarization. The end point of the T wave referred to as T_{off} and represents the end of the cardiac cycle (presuming the absence of a U wave). The U-wave is timely related to the diastolic phase of the cardiac mechanical cycle since it may start immediately before the second heart sound at the onset of the ventricular relaxation. Major theories of origin and physiology of the U-wave are summarized in [14].

The PR interval extends from the start of the P wave to the end of the PQ junction at the very start of the QRS complex. Therefore, this interval is sometimes known as the PQ interval, representing the time required for the electrical impulse to travel from the sinoatrial node to the ventricle and normal values range between 120 and 200 ms. The global point of reference for the ECG's amplitude is the isoelectric level, measured over the brief period between the P wave and QRS complex. Since there is a short pause before the current is conducted between the atria and the ventricles, this point is generally thought to be the most stable marker of 0V for the surface-ECG. The QRS width is representative of the time for the ventricles to depolarize, typically lasting 80 to 120 ms. The lower the heart rate, the wider the QRS complex is, due to decreases in conduction speed through the ventricle. The RS segment of the QRS complex is known as the ventricular activation time and is usually shorter (lasting around 40 ms) than the QR segment. This asymmetry in the QRS complex is not a constant and varies based upon changes in the autonomic nervous system, lead position, respiration, and heart rate. The QRS complex usually rises (for positive leads) or falls to about 1 to 2 mV from the isoelectric line for normal beats. The QT interval is considered to represent the time between the start of ventricular depolarization and the end of ventricular repolarization, thus it can be used as a measure of the duration of repolarization [7]. Typical ECG features and their normal values in a sinus rhythm of a healthy male adult, specifically considering lead II, with 60 bpm as heart rate are reported in detail in *Table I*.

Table I – Typical Lead II ECG features and their normal values in sinus rhythm of a healthy male adult having 60 bpm as heart rate. From [7].

<i>Feature</i>	<i>Normal value</i>	<i>Normal limit</i>
<i>P width</i>	110 ms	±20 ms
<i>PQ/PR interval</i>	160 ms	±40 ms
<i>QRS width</i>	100 ms	±20 ms
<i>P amplitude</i>	0.15 mV	±0.05 mV
<i>QRS height</i>	1.5 mV	±0.5 mV
<i>ST level</i>	0 mV	±0.1 mV
<i>T amplitude</i>	0.3 mV	±0.2 mV

The ECG is the most widely used initial diagnostic test, able to provide vital information about cardiac rhythm, presence of arrhythmias, myocardial ischemia/infarction, and other kind of

abnormalities. To date, the 12-lead ECG remains the gold standard used for initial screening, identifying, and evaluating patients with chest pain and angina pectoris. This recording modality may be derived from the orthogonal Frank lead configuration by the inverse Dower transform [15]. An electrode defined as a conductive pad attached to the skin that records heart signals by converting ionic signals from the body into electrical signals. The specific standard uses 10 electrodes to record the electrical activity of the heart: four limb leads (Right Arm (RA), Right Limb (RL), Left Arm (LA), Left Limb (LL)) and six precordial (chest) leads (V1, V2, V3, V4, V5, and V6). As shown in Fig. 5, the six precordial leads, indicated in red, are placed across the precordium in anatomically specific locations and four limb leads, shown in green, may be placed either on the distal limbs, the preferred placement for standard resting 12-lead ECG acquisition, or alternatively where the limbs attach to the torso (Mason-Likar) for continuous ECG monitoring, as required in an exercise testing [16]. The correct placement of the chest and limb electrodes described in Table II.

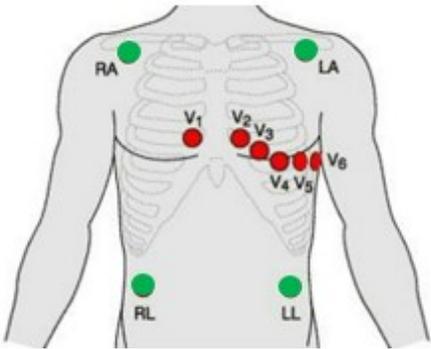


Fig. 5 – The standard 12-lead ECG. Red dots: precordial (chest) leads. Green dots: limb leads.

Table II – Chest and Limb electrodes placement according to the standard 12-lead ECG. Right Arm (RA), Right Limb (RL), Left Arm (LA), Left Limb (LL).

Chest electrodes	V1	Fourth intercostal space on the right sternum
	V2	Fourth intercostal space on the left sternum
	V3	Midway between placement of V2 and V4
	V4	Fifth intercostal space at the midclavicular line
	V5	Anterior axillary line on the same horizontal level as V4
	V6	Mid-axillary line on the same horizontal level as V4 and V5
Limb electrodes	RA	Anywhere between the right shoulder and right elbow
	RL	Anywhere below the right torso and above the right ankle
	LA	Anywhere between the left shoulder and left elbow
	LL	Anywhere below the left torso and above the left ankle

The six precordial leads record heart activity in the *horizontal* plane. These are unipolar leads and consist of a single positive electrode (exploring electrode) with a reference point found at the electrical center of the heart, known as Wilson central terminal (WCT), a virtual point placed at the center of the chest or the center of Einthoven's triangle. WCT is an average of the potentials from the limb leads computed by connecting all these leads to one terminal via electrical resistance. Anatomically the ECG leads V1 and V2 look at the right ventricle, V3 and V4 at the septum (stout wall) between the ventricles and the anterior wall of the left ventricle, V5 and V6 at the anterior and lateral walls of the left ventricle. Potential differences between the limb electrodes and the central terminal are the bases for the other three standard ECGs: (1) Lead I, the difference between LA and RA; (2) Lead II, the difference between the LL and the RA; (3) Lead III, the difference between the LL and the LA. From that basic configuration, it is also possible to obtain additional useful information: aVF as the difference between the LL and average of the arm leads; aVR as the difference between the RA and the average of LL and LA; aVL as the difference between the LA and the average of the RA and LL [7].

1.4. Myocardial Ischemia: pathophysiology, clinical presentation, and underlying mechanisms

The most common condition among those included in the broader category of IHD is the myocardial ischemia, described in detail in this section.

Myocardial ischemia, a temporary insufficient perfusion of a particular region in the heart, occurs when there is a mismatch of blood supply and demand due to insufficient supply given the demand (supply ischemia) or more demand than the available supply (demand ischemia), causing cardiac dysfunction, arrhythmias, myocardial infarction, and sudden death. The ability of the supply chain to meet the increased demand depends on the ability of the coronary circulation to dilate and increase the flow as required. The insufficiency in blood supply may be caused by ACS, coronary spasm, or anemia. The most sensitive area to ischemia and the first to suffer oxygen deficiency is the subendocardium, leading to subendocardial ischemia. When the entire myocardium is affected by the delayed repolarization, the disease evolves into subepicardial or transmural ischemia.

The identification of the mechanism(s) that induce ischemia in each patient appears of paramount importance for a tailored therapeutic approach. The new perception of myocardial ischemia as a

multifactorial condition, proposed by [17], expected to downgrade the role of coronary stenosis, and in greater attention to alternative mechanisms of ischemia. Pathophysiologic mechanisms, interacting through intricate feedback pathways and in a patient-specific manner, are reported below.

1. Atherosclerosis

Atherosclerosis, a chronic disease, may morphologically be present in coronary arteries as either eccentric or concentric lesions. It may have a 2-fold effect on CBF: in the first place, it may be responsible for coronary artery stenosis causing an increase in resistance and therefore a reduction in coronary flow. In patients with atherosclerosis, an increase in myocardial oxygen consumption may not be matched by a sufficient increase in coronary flow. The resultant myocardial ischemia is therefore more likely to originate from the inconsistencies in coronary flow reserve rather than variability of myocardial oxygen consumption [18].

2. Coronary Artery Spasm

Although rare, the isolated coronary spasm is a further distinct cause for myocardial ischemia, otherwise known as Prinzmetal or variant angina, as discussed by several reports, as [19] [20]. This syndrome usually occurs at rest and caused by vasospasm in (often normal) coronary arteries rather than atherosclerosis, but also at the focal site of an atherosclerotic plaque of variable severity. The pathophysiology of coronary artery spasm is multifactorial, but the most important causative factor is an increased intracellular calcium concentration in combination with elevated calcium sensitivity [18].

3. Thrombus formation

A common cause of ischemia associated with a decrease in blood supply is the occurrence of an adverse event in an epicardial coronary artery, such as the rupture, ulceration, fissure, or erosion of an atherosclerotic plaque, that introduces thrombogenic factors into the bloodstream, resulting in intraluminal thrombus formation and, consequently, partial or complete occlusion of the blood vessel. When the thrombus does not resolve spontaneously, and there is no timely effective interventions, such those reported in the following *Section 2.1*, a myocardial infarction develops within hours [21].

4. Vasomotor impairment

Vasomotor impairment characterized by enhanced reactivity of vascular smooth muscle, resulting in spasm or impaired vasodilation of the microvascular and/or epicardial compartments. Up to one-third of patients with ischemia and non-obstructive coronary

diseases present vasomotion disorders. It can be endothelium-independent, reflecting the coronary flow reserve's inability to adapt to the vasodilatory stimuli, or endothelium-dependent with a paradoxical response at Acetylcholine (Ach) administration. These conditions are a frequent underdiagnosed cause of ischemia and are particularly prevalent in women [22].

Metabolic changes, including myocardial lactate production, coronary sinus oxygen desaturation, and pH reduction in the coronary sinus, are also important objective proof of myocardial ischemia. In relation to this, myocardial release of lipid peroxide products in the coronary circulation is a marker of myocardial ischemia with a high sensitivity even for brief and/or mild myocardial ischemia [23]. The degree of the coronary occlusion, the volume of the affected ischemic myocardium, extent of collateral circulation, pre-existing myocardial metabolic rate, genetic factors, and the intrinsic survival capacities of the myocytes will collectively determine the specific clinical presentation and the severity of myocardial ischemia [18]. The reduced blood supply causes depolarization of resting membrane potential of the ischemic region with respect to the resting membrane potential of the normal region, resulting in ECG changes, especially ST-segment elevation/depression (STE/STD) and T wave alterations [13].

Chronic coronary syndrome comprises chronic coronary atherosclerotic lesions and coronary microvascular dysfunction (syndrome X). Ischemia in this case is usually caused by exercise, but it may also occur in rest because of increased demand due to tachycardia, anemia, or a rise in blood pressure [21].

Symptoms suggestive of cardiac ischemia include retrosternal chest pain (with or without radiation to either arm, the neck, or the jaw), oppressive chest pressure, abdominal pain, dyspnea, nausea, vomiting, diaphoresis, and syncope [24].

When the partial or complete occlusion of the blood vessel by the thrombus occurs in the coronary circulation, this scenario called ACS [25], as described detailly in the next paragraph.

1.5. Acute Coronary Syndrome: pathophysiology, clinical presentation, and underlying mechanisms

ACS is a potentially life-threatening condition that affects millions of individuals each year, remaining an important public health concern. Over the past years, studies have led to an improved

understanding of the pathophysiology of ACS and advancements have made in the medical management of this condition. ACS describes the range of myocardial ischemic states that includes unstable angina (UA), non-ST elevated myocardial infarction (NSTEMI; often referred to as “non-Q-wave myocardial infarction”), or ST-elevated myocardial infarction (STEMI; often referred to as “Q-wave myocardial infarction”). The diagnosis and classification of ACS is based on a thorough review of clinical features, including ECG findings and biochemical markers of myocardial necrosis [26].

ACS are rarely caused by coronary dissection, arteritis, myocardial bridging, thromboembolism, or coronary vasospasm without obvious coronary artery disease (CAD). With very few exceptions, coronary atherosclerosis is the underlying condition for ACS, consisting of the ongoing process of plaque formation that involves primarily the intima of large- and medium-sized arteries. The formation of occlusive blood clot on a ruptured atherosclerotic plaque in an epicardial coronary artery causes an acute reduction of blood flow and a complete or almost complete obliteration of the coronary artery lumen. If reperfusion does not occur rapidly, this conditions worsens in the acute transmural ischemia, that involves all layers of the myocardium with necrosis of the involved tissue [27]. The subsequent four mechanisms, summarized in *Fig. 6*, are responsible for ACS:

A. Plaque fissure with inflammation

Plaque rupture, also referred to as fissure, traditionally considered the dominant substrate for ACS and is typically associated with both local and systematic inflammations. The latter is often indicated by an increase in blood C-reactive protein (CRP), which can be measured using a high-sensitivity (hsCRP) assay. Plaque rupture defined as a structural defect, a gap, in the fibrous cap that separates the lipid-rich necrotic core of a plaque from the lumen of the artery. This condition causes the growing of a “red” thrombus, so called by the erythrocyte-rich. Several laboratory studies and observations have demonstrated that inflammatory mechanisms are key regulators of the fragility of the fibrous cap and of the thrombogenic potential of the lipid core. An increased amounts of activated macrophages with degrading-related function, or reduced levels of their corresponding endogenous inhibitors, can enhance breakdown of the arterial extracellular matrix of plaque.

B. Plaque fissure without inflammation

In some cases, plaque rupture complicates atheroma that is not associated with elevations in circulating CRP and with a feeble systemic inflammatory activation. Plaque rupture usually causes the formation of fibrin-rich “red” thrombi. Intense physical exertion and local mechanical stress at the level of the artery wall, either increased circumferential stress or

reduced shear stress, may predispose to that pathogenesis. Joined in these, extreme emotional disturbance ranging from external events of short duration (as earthquakes) to acute manifestations of more long-term emotional disturbances may contribute to plaque rupture.

C. Plaque erosion

Plaque erosion starts with changes in endothelial shear stress gradients, resulting in loss of basement membrane integrity and endothelial cell desquamation. The thrombi, overlying patches of intimal erosion, exhibit characteristics of white platelet-rich structures, also named “white” thrombi. The neutrophil activation, because of attraction by chemokines produced by activated endothelial cells, plays a pivotal role in thrombosis due to plaque erosion, more commonly associated with hypertriglyceridemia, diabetes mellitus, female sex, and old age. It has become the predominant mechanism of NSTEMI.

D. Vasospasm

Vasospasm is long recognized as a phenomenon mainly in the epicardial arteries but also affecting coronary microcirculation. It is a condition in which the muscular walls of an artery suddenly contract, causing the artery to narrow and reduce the amount of blood that can flow through it. Either macrovascular or microvascular spasm can result from impaired vasodilatation or from vasoconstrictor stimuli acting on hyperreactive vascular smooth muscle cells [28].

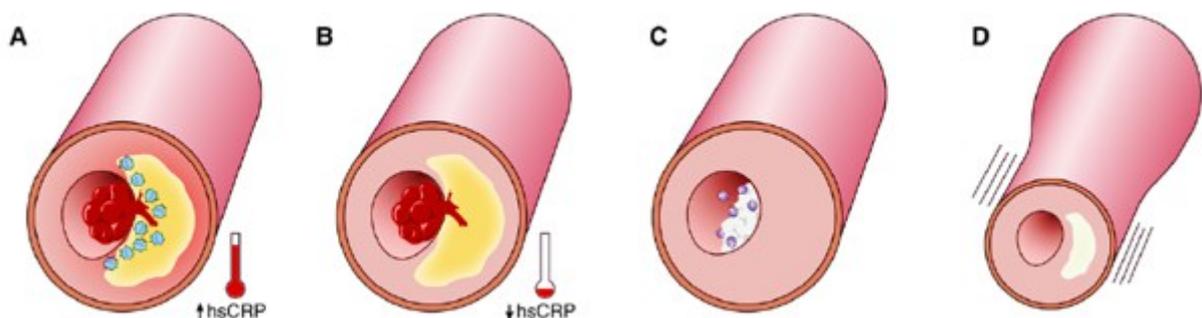


Fig. 6 – Four mechanisms causing ACS. (A) Plaque fissure with inflammation, where the red mass indicates the thrombus and the blue dots the macrophages reaching lesions. (B) Plaque fissure without inflammation, characterized by low systematic inflammation. (C) Plaque erosion characterized by “white” thrombus and the presence of neutrophils, but without fissure. (D) ACS without thrombus, in which an epicardial or microvascular spasm occurs. From [28].

The ratio between smooth muscle cells and macrophages, derived from the monocyte differentiation after endothelium damage, plays an important role in plaque vulnerability and the tendency to rupture. Although plaque rupture results in ACS, in 99% of cases, it is clinically silent.

The rate of progression of atherosclerotic lesions is variable, nonlinear, and unpredictable [29]. Dynamic changes in the size of the thrombus with distal embolization of platelet aggregates and clots and secretion of vasoactive substances lead to cyclic flow variations with repeat episodes of subendocardial ischemia, so-called unstable angina, which may lead to myocardial cellular injury, named acute myocardial infarction [27].

Unstable angina and NSTEMI are closely related conditions: their pathophysiologic origins and clinical presentations are similar, but they differ in severity. A diagnosis of NSTEMI can be made when the ischemia is sufficiently severe to cause myocardial damage that results in the release of a biomarker of myocardial necrosis into the circulation (cardiac-specific troponins T or I, or muscle and brain fraction of creatine kinase [CK-MB]). In contrast, the patient considered to have experienced UA if no such biomarker can be detected in the bloodstream hours after the initial onset of ischemic chest pain. Unstable angina exhibits 1 or more of 3 principal presentations: (1) rest angina (usually lasting >20 minutes), (2) new-onset (<2 months previously) severe angina, and (3) a crescendo pattern of occurrence in terms of intensity, duration, frequency [29].

By 6 months, UA/NSTEMI mortality rates higher than that after STEMI can be seen; and by 12 months, the rates of death, myocardial infarction, and recurrent instability in contemporary randomized controlled trials and registry studies exceed 10% and are often related to specific risk factors such as age, diabetes mellitus, renal failure, and impairment of left ventricular (LV) function [30].

From a review of the literature [31], it was found that the vast majority of coronary thrombi (73%) developed on top of a ruptured atherosclerotic plaque. Plaque rupture is a commoner cause of coronary thrombosis in men (80%) than in women (60%) and it is especially rare in premenopausal women [3]. The condition progresses uninterruptedly throughout a person's lifetime, before finally manifesting as an acute ischemic event. Several coronary risk factors influence this process, including hypercholesterolemia, hypertension, diabetes, and smoking. These damage the endothelium of the blood vessel, playing a pivotal role in initiating the pathological process. Dysfunctional endothelium impairs vascular hemostasis, due to the reduced bioavailability of nitric oxide and the excessive production of endothelin. Additionally, in this area the increase of the adhesion molecules expression facilitates the activation of monocytes and enhances the thrombogenicity of blood through the secretion of several locally active substances [29].

A diagnosis of ACS should be considered in all patients presenting with ischemic symptoms. Symptoms of ACS include chest pain, referred pain, nausea, vomiting, dyspnea, diaphoresis, and

lightheadedness. Pain may be referred to either arm, jaw, neck, the back, or even the abdomen. Typical angina described as pain that is substernal, occurs on exertion, and is relieved with rest. Patients with all three of these features have a greater likelihood of having ACS than patients with none, one, or even two of these features [32]. The pain and discomfort associated with an ACS event may occur with exertion or at rest and is often diffuse rather than localized, as in ischemia symptomatology [29]. Ominous physical findings include a new mitral regurgitation murmur, hypotension, pulmonary rales, a new third heart sound (S3 gallop), and new jugular venous distention [32].

Chapter 2 – Triage & Assessment of syndromes

2.1 Triage

Cardiovascular emergencies account for approximately 10% of all ED visits [33]. More than 8 million patients with chest pain and/or anginal equivalent symptoms present to ED each year, accounting for the second, most common cause of ED visits for adults. The ED is a fast-paced, dynamic, and chaotic setting that requires quick and accurate decision making to distinguish high-acuity patients. The priority for emergency clinicians is to recognize and stabilize patients with emergent cardiovascular conditions that include, but are not limited to, myocardial ischemia/infarction and potentially life-threatening arrhythmias [16]. When a patient presents chest pain or symptoms suggestive of ACS, vital signs should be obtained, monitored and a careful history should be obtained. Accurate risk stratification and diagnostic testing are critical for time dependent therapies that restore blood flow to the compromised myocardium [34]. In contrast, inaccurate triage could result in flooding of emergency/cardiology departments, performing unnecessary urgent catheterization and/or administering potentially hazardous thrombolytics, while false-negative cases would miss important treatment. Individuals experiencing acute chest pain in the community represent an undifferentiated population, often presenting ad hoc to first medical responders in the pre-hospital setting. In the Multicenter Chest Pain Study [35], acute ischemia was diagnosed in 22% of patients who presented to the ED with sharp or stabbing pain and in 13% of patients with pain with pleuritic qualities. Furthermore, 7% of patients whose pain was fully reproduced with palpation were ultimately recognized to have ACS. These patients should undergo immediate risk assessment and triage following local protocols established within the emergency medical service.

The two primary approaches for optimizing patient outcomes are to reduce treatment delay and to minimize total ischemic time, defined as time of symptom onset to reperfusion of culprit arteries. To minimize total ischemic burden time, the American Heart Association and European Society of Cardiology (AHA/ESC) recommend that individuals with chest pain seek medical attention immediately and receive a 12-lead ECG within 10 min of hospital arrival, based on evidence that longer delays are associated with adverse prognoses [36]. Additionally, the initial ECGs has to be repeated at 5- to 10-minute intervals if the initial ECG is not diagnostic, but the patient remains symptomatic and a high clinical suspicion for ACS persists [16]. Prolonged chest pain (>15 min) and/or recurrent pain within 1h are the primary symptoms that prompt consideration of the clinical

diagnosis of ACS and the initiation of testing according to specific diagnostic algorithms. To understand the complexity of ACS-related symptomatology, careful history taking and comprehensive interaction with the patient are crucial for achieving an early and accurate diagnosis. The resting 12-lead ECG is the first-line diagnostic tool in the assessment of patients with suspected ACS and ischemic symptoms. After this, patients can be further classified based on the presence or absence of biomarkers (once these results are available), testing over a 6- to 12-hour period. These features are important in the initial triage and diagnosis of patients with ACS, to risk stratify patients and guide the initial management strategy [4].

Measurement of a biomarker of cardiomyocyte injury, preferably high-sensitivity cardiac troponin (hs-cTn), recommended in all patients with suspected ACS, although measurable levels in the bloodstream reached 3 to 6 h after the damage. Among the multitude of biomarkers evaluated for the diagnosis of NSTEMI, only creatine kinase myocardial band isoenzyme, myosin-binding protein C, and copeptin may have clinical relevance when used in combination with cTn T/I, although in most clinical situations their incremental value above and beyond cTn is limited. Ischemia-modified albumin (IMA) has been demonstrated to be a biomarker of ischemia associated with myocardial and skeletal muscle ischemia, pulmonary embolism, and stroke. Blood levels of IMA rise quickly after the onset of ischemia, within 5 to 10 min, and continue to rise while the condition persists. When combining patients' signs and symptoms with ECG results and cardiac-specific troponin levels, the diagnostic accuracy is approximately 50%. With the addition of the recent FDA approved ischemia-modified albumin assay, the diagnostic accuracy has increased to approximately 70% [9]. Focused physical examination should include checking for the presence of all major pulses, measurement of blood pressure in both arms, auscultation of the heart and lungs, and assessing for signs of heart failure or circulatory compromise [4].

Depending on the initial assessment of the ECG, the clinical context and hemodynamic stability, patients with suspected ACS should be classified as either:

- A. Patients with a working diagnosis of STEMI are triaged for immediate reperfusion therapy, undergoing to the primary percutaneous coronary intervention (PPCI) strategy or fibrinolysis, if PPCI is not possible within 120 min of diagnosis. A PPCI intends an emergent PCI with balloon, stent, or other approved device, performed on the Infarct-Related Artery (IRA) without previous fibrinolytic treatment. Whereas fibrinolytic therapy is an important reperfusion strategy for STEMI patients, preventing 30 early deaths per 1000 patients treated within 6 h of symptom onset. Acute fibrinolytic therapy is contraindicated for ACS patients

without STE, except for those with electrocardiographic true posterior myocardial infarction manifested as STD in two contiguous anterior precordial leads and/or isolated STE in posterior chest leads [30]. Successful reperfusion is generally associated with significant improvement in ischemic symptoms, $\geq 50\%$ ST-segment resolution, and hemodynamic stability. Emergency coronary artery bypass grafting (CABG) surgery should be considered for patients with a evident IRA, but with unsuitable anatomy for PCI, and either a large myocardial area at jeopardy or with cardiogenic shock [4].

- B. For patients with a confirmed or working diagnosis of NSTEMI, a routine invasive strategy with inpatient coronary angiography recommended for reducing the risk of composite ischemic endpoints, paying the risk of periprocedural complications and bleeding. They should undergo pre-hospital triage in accordance with protocols for patients in the STEMI pathway, since they also face immediate risks, including ventricular arrhythmias. Conversely, for patients with a low index of suspicion, a selective invasive approach recommended [4].

Moreover, the ECG considered the key objective prehospital diagnostic, therapeutic, and prognostic tool also for myocardial ischemia detection. According to the guidelines [25], the ECG evaluated for signs of STE or STD measured at the J-point. While J-amplitude deviations frequently occur alongside myocardial ischemia, the alterations in myocardial action potentials due to ischemia generate injury currents throughout all phases of the cardiac cycle. This results in observable ECG changes throughout the entire QRST complex [37].

While STE presents the greatest early risk of mortality, STD observed on the initial ECG is associated with the highest risk of death at 6 months and its severity demonstrates a strong correlation with patient outcome. Because patients with ischemic discomfort at rest are heterogeneous in terms of risk of cardiac death and nonfatal ischemic events, an assessment of the prognosis guides the initial evaluation and treatment. An estimation of risk is useful in: (1) selection of the site of care (coronary care unit, monitored step-down unit, or outpatient setting); (2) selection of therapy, including platelet glycoprotein (GP) IIb/IIIa inhibitors and invasive management strategy. An immediate invasive strategy, referred to angiography and PCI, is recommended for patients with a recurrent dynamic ECG change suggestive of ischemia (particularly with intermittent STE) [38]. Regardless of the angiographic strategy, an assessment of LV function is recommended in patients with documented ischemia because of the imperative to treat patients who have impaired LV function with Angiotensin-Converting Enzyme (ACE) inhibitors, beta blockers, and, when heart failure or diabetes mellitus is present, aldosterone

antagonists; instead, when the coronary anatomy is appropriate (3-vessel coronary disease), CABG is done [30].

After the acute and stabilization phases, most aspects of the subsequent management strategy are common to all patients with ACS and myocardial ischemia (regardless of the initial ECG pattern or the presence/absence of biomarker elevation at presentation) and can therefore be considered under a common pathway. Revascularization will often reverse both the T-wave inversion and wall-motion disorder. Antithrombotic treatment is an important component of the management of all patients presenting with suspected symptoms. The specific choice and combination of therapy, the time of its initiation, and the treatment duration depend on various patient and procedural factors. Treatment decisions must be made to weigh the benefits of the therapy against the risk of bleeding. ECG monitoring for arrhythmias and new STE/STD recommended for at least 24 h after symptom onset in all high-risk patients [4]. At the end of the observation period, the patient is reevaluated and then generally undergoes functional cardiac testing (resting nuclear scan or echocardiography) and/or stress testing (treadmill, stress echocardiography, or stress nuclear testing) or noninvasive coronary imaging study (coronary computed tomographic angiogram (CCTA)). Those patients who have a recurrence of chest pain strongly suggestive of ACS, a positive biomarker value, a significant ECG change, or a positive functional/stress test or CCTA, are generally admitted for inpatient evaluation and treatment. Patients with continuing discomfort and/or hemodynamic instability should be hospitalized for at least 24 h in a coronary care unit, undergoing continuous ECG monitoring during their ED evaluation and early hospital phase, because sudden, unexpected ventricular fibrillation is the major preventable cause of death in this preliminary period. The American College of Emergency Physician (ACEP) has published guidelines that recommend a program for the continuous monitoring of outcomes of patients evaluated in such units and the impact on hospital resources. The acute phase of UA/NSTEMI is usually over within 2 months. The risk of progression to myocardial infarction or the development of recurrent myocardial infarction or death is highest during that period. At 1 to 3 months after the acute phase, most patients resume a clinical course similar to those with chronic stable coronary disease. Patients who have undergone successful PCI with an uncomplicated course are usually discharged the next day, and those with an uncomplicated CABG 4 to 7 days after the surgery. Other less extensively studied therapies for the relief of ischemia but not been applied in the acute setting for UA/NSTEMI, such as spinal cord stimulation and prolonged external counter pulsation, are under evaluation [30].

2.2 Electrocardiography as a diagnostic tool

Known that the extended ischemia may lead to cardiac cell death, and it is a pre-condition to infarction, the identification of myocardial ischemia, as well as ACS, at the earliest phase is substantial to ban the destructive results. A meticulous evaluation of ECG changes can allow in estimating time of the event, amount of myocardium at risk, patient prognosis, and appropriate therapeutic strategies, thereby reducing morbidity as well as mortality and avoiding inappropriate hospital discharge. Specifically, the standard 12-lead ECG is the common criterion for electrocardiographic evaluation of acute myocardial ischemia/injury and is the most immediately available diagnostic modality for patients with suspected ACS in the ED. While ECG value for the assessment of ACS, it is not overly sensitive or specific for the detection of myocardial ischemia [39].

Changes in the intracellular action potential during myocardial ischemia, injury, and infarction result in changes in ECG waveforms, which reflect the alterations affecting the main cardiac vector. In fact, during an ischemic episode, the reduced flow caused by coronary occlusion leads to regional metabolic derangements, which distort the morphology of the action potential and the modality of excitation propagation throughout various myocardial segments. ECG changes indicating ischemia include STE, STD, or T-wave inversion and occur before myocardial infarction, providing the ability to restore blood flow before myocardial cell death ensues. The ST-segment shifts are determined by the flow of currents, referred to as “injury currents”, generated by the voltage gradients across the boundary between the ischemic and nonischemic myocardium, during the resting and plateau phases of the ventricular action potential, which correspond to the TQ and ST segments of the ECG [40]. In the ECG recorded with electrode facing the ischemic zone, the fall in resting potential during the first minutes is reflected in an initial depression of the TQ segment, classically seen as a STE. Transiently peaked T waves with lengthening of the QT interval are, indeed, the first manifestations of acute myocardial ischemia in case of sudden complete occlusion of an epicardial coronary artery, including coronary spasm. The intrinsic deflection delayed and the QRS broadened. After 5 min, the ST segment becomes further elevated because of the shorter action potential in the ischemic part of the heart, resulting in transmural ischemia. Later, when activation in the ischemic region is seriously delayed, the ST segment becomes markedly elevated and followed by a pronounced inverted T wave, usually with the development of pathological Q waves depending on the resulting amount of necrosis. After 8–15 min, local activation transiently recovers and is often accompanied by T-wave alternans [40] [41].

T wave evolution in IHD is not a marker of cell death, instead caused by changes in the ion channels in regions of the heart that remain viable after an episode of severe ischemia. T-wave inversion is sensitive to ischemia but is less specific, unless it is marked (≥ 0.3 mV) [29]. Unfortunately, signs of ischemia in the QRS-complex and in the T wave cannot readily be detected due to the wide ranges of normal values which can overlap with ischemic changes. Additionally, nonacute pathology, such as left ventricular aneurysm, can severely alter the ECG [37].

Since only small deviations from the isoelectric level are significant markers of cardiac abnormality, the correct measurement of the isoelectric line is crucial. Normally, J-point expected to be isoelectric, as it is the pause between ventricular depolarization and repolarization. The ST level is generally measured around 60 to 80 ms after the J-point, with adjustments for local heart rates [16]. Abnormal changes in the ECG, defined by the Sheffield criteria, described in [42], are ST level shifts ≥ 0.1 mV (or about 5% to 10% of the QRS amplitude for a sinus beat on a V5 lead). Whether the ST segment is elevated or depressed, it depends on the positions of the recording electrodes with respect to the heart. Several patterns of STD can be seen: horizontal, down-sloping or up-sloping and accompanied by tall positive, biphasic or negative T waves. When ischemia confined primarily to the subendocardium, the overall ST vector typically faces the inner ventricular layer and the ventricular cavity such that the surface ECG leads show horizontal or down-sloping STD. This subendocardial ischemic pattern is a frequent finding during spontaneous episodes of angina at rest and represents the typical ECG finding during exercise tests. Typically, maximal STD in demand ischemia recorded in the precordial leads V4-V6. In cases of severe extensive subendocardial ischemia, as in acute subtotal or even total occlusion of the left main coronary artery, the injury vector may be seen as STD in most of the ECG leads, while STE in lead aVR. In these cases, extensive ischemia impairs relaxation of the left ventricle [40] [43]. The ST Segment Monitoring Practice Guideline Working Group, discussed in-depth in [44], recommends that if only two leads are available for ST segment monitoring, leads III and V3 should be used. Instead, the best three-lead combination is III-V3-V5. In addition, these two- and three-lead combinations for ischemia exclude lead V1, considered the best lead to monitor for detection of cardiac arrhythmias. Furthermore, the use of at least three chest leads (V3, V4, V5) recommended for ST analysis, to allow noise reduction and artifact identification (although four- or five-lead configurations give better results) [7]. For instance, an ischemic event, occurred in an adult male with transient chest pain thought to be due to coronary vasospasm, can be evidenced in ECG waveforms recorded in the leads I, II, V1, V3 and V6, as shown in *Fig. 7*.

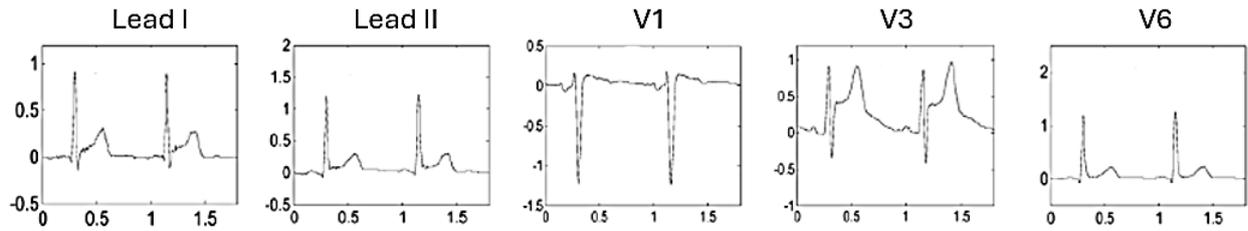


Fig. 7 – Electrocardiographic waveforms in lead I, II, V1, V3 and V6 during an ischemic event occurred in an adult male with transient chest pain. From [45].

By using ST-T segment deviations recorded at body surface, even of only 0.05 mV, estimation for transmembrane potential through heart wall was capable to identify the position and size of ischemic region. The sensitivity and positive predictivity of ST segment in ECG, found by [13], are of an average of 87-89% and 90-92%, respectively. Concomitant ischemia of opposing segments may lead to cancellation of ST deviation in leads facing these segments, since various ECG leads record the vector of the global activation of the heart toward and away of the electrode and not local events [27].

Several factors will affect the ECG pattern observed in an individual patient with ACS. The most important are: 1) the cellular consequences induced by myocardial ischemia from total coronary artery occlusion versus subtotal occlusion with or without distal embolization; 2) the duration/extension/severity/localization of the ischemic process; 3) the presence of underlying abnormalities, such as intraventricular conduction disturbances, repolarization abnormalities, pacemaker implant; 4) individual variation in coronary anatomy [40].

The diagnosis of ACS includes both STEMI, which phases are shown in *Table III*, and NSTEMI. The STEMI ECG criteria are concrete, qualitative, and quantitative. A STEMI ECG pattern is defined as an STE at the J-point of ≥ 0.2 mV (0.15 mV in women; 0.25 mV in men <40 years) in two or more contiguous leads in leads V1, V2 or V3, and of ≥ 0.1 mV in other contiguous leads (contiguity in the frontal plane is defined by the lead sequence aVL, I, inverted aVR, II, aVF and III). Detection of ischemia by the STEMI criteria can fail because the ST injury vector is too small, even in complete occlusions, or assumes a direction that is not optimal for projection on the lead vectors of the involved leads [21].

Table III – Electrocardiographic phases of STEMI.

<i>Phase 0</i>	<i>Phase 1</i>	<i>Phase 2</i>	<i>Phase 3</i>	<i>Phase 4</i>
Peaking-T: positive, high, wide T wave	Q small, R small, monophasic STE, T positive	Q large, R small, STE regressing, T peaked and negative	Q large, R increase, STE vanished	Q still large, R normal dimension, no STE, no ST- decline, T positive again
				

Patients may present with tall positive T waves without ST deviation or with minimal ST deviation. It has been suggested that persistent positive T waves without STE can be caused by an occlusion of an artery with preexisting severe narrowing and well-developed collateral circulation. Some of these patients will develop abnormal Q waves without significant STE, but the majority will progress to STEMI followed by Q-wave infarction. Differently, the presence of abnormal Q waves in leads with STE may signify a more advanced stage of infarction with less potential for myocardial salvage by reperfusion therapy and poorer prognosis, irrespective of whether they present necrosis or local conduction delay [43]. Established Q waves greater than or equal to 0.04s are also less helpful in the diagnosis of UA, although by suggesting prior myocardial infarction, they do indicate a high likelihood of significant CAD. Isolated Q waves in lead III may be a normal finding, especially in the absence of repolarization abnormalities in any of the inferior leads [30].

In contrast, the ECG pattern associated with NSTEMI is not localized. Any ECG not complying with the STEMI ECG criteria belongs to that class of ECGs, if they recorded in an ACS patient. NSTEMI ECGs are not extremely specific for the presence of ischemia: >30% of the ECGs in this patient group are normal, and STD and T-wave inversion can also be seen in patients with other pathologies. ACS-related ischemic conditions expressing as an NSTEMI ECG pattern vary from mild to severe, the latter in left main or multivessel disease or left circumflex occlusions. NSTEMI includes ECGs with STD but also without ST-segment changes, the latter for various reasons, as summarized in [21]. Several distinct ECG patterns seen in patients with NSTEMI have been characterized: 1) T-wave inversion with isoelectric ST segments or minor ST deviation; 2) Up-sloping STD with positive tall T waves; 3) Diffuse STD in the inferior and anterolateral leads associated with STE in lead aVR [27]. One group of investigators [46] found that the diagnosis of

NSTEMI is greater than three times more likely in patients with chest pain whose ECG showed STD in three or more leads or STDs that were greater than or equal to 0.2 mV. This finding supported by [32], which concluded that approximately 25% of patients with STD eventually develop STEMI, and the remaining 75% percent have NSTEMI. Findings on ECG associated with UA include STD, transient STE, T-wave inversion, or a combination of these factors; depending on the severity of the clinical presentation, are common in 30% to 50% of patients [29].

In addition to the changes reflecting acute ischemia and reperfusion, the ECG gives information concerning the presence of preexisting coronary heart disease and hence, myocardial reserves. Changes related to “footprints” of the preceding ischemia (minor ST deviation, terminal, or complete T-wave inversion) or signs of evolved myocardial infarction (STE with terminal T-wave inversion, Q waves) can be found.

- 1) In the absence of symptoms, STE may be secondary to nonischemic causes or reflect residual changes of the previous ischemia, including aneurysm.
- 2) In the absence of symptoms, STD may reflect residual effects of the preceding ischemia. However, in some cases it is caused by chronic preexisting conditions, as left ventricular hypertrophy with secondary repolarization changes and cardiomyopathy.
- 3) In the ACS setting, T-wave inversion without concomitant STD is a sign of reperfusion of prior active ischemia, not of ongoing active ischemia itself.
- 4) Q waves in leads without acute changes indicate prior myocardial infarction/scar [43].

2.3 Serial electrocardiography

Although ECG is an essential part of the initial evaluation of patients with symptoms suspected to be related to myocardial ischemia and ACS, the complex nature of chest pain combined with unpredictable and dynamic acute ischemic changes suggests that a single 12-lead ECG snapshot is not always sufficient to make an accurate diagnosis. A serial approach, consisting of comparing the current ECG to a previously acquired ECG, corrects for interindividual variability, reveals the actual intra-individual ischemic changes, and is, indeed, recommended by the guidelines [25]. The availability of prior ECG tracing and repeated ECG of the same patient could enhance the diagnostic performance and timeliness for initiating medical treatment. Moreover, ECGs have a wide range of normal morphologies, and it is well possible that a person with a preexisting ECG that is within normal limits will still be within normal limits when ischemic changes occur. This

might be an explanation why, part of the NSTEMI patients have a normal prehospital ECG, and an ischemic ECG change remains occult. On the other hand, if the preexisting ECG of the subject is not within normal limits, it may during ischemia also be challenging to discover the ischemic component in the ECG. Likely, the best way to measure ischemic ECG modification in either of these two examples is to do a differential measurement, serially comparing the preexisting and the acute ECG. The comparison of an acute, ischemic ECG, with a previously electively recorded nonischemic ECG is subject to several factors that might interfere with the concept that the differences between the current and the historical ECG are effectively diagnostic [21]. For instance, looking at *Fig. 8* from [47], the diagnosis of a preexisted left-ventricular hypertrophy, shown on the left panel, makes challenging to recognize the ischemic component in the right panel, if taken and interpreted alone, while a serial comparison makes these ischemic changes in the QRS complex, the ST segment, and the T wave readily apparent.

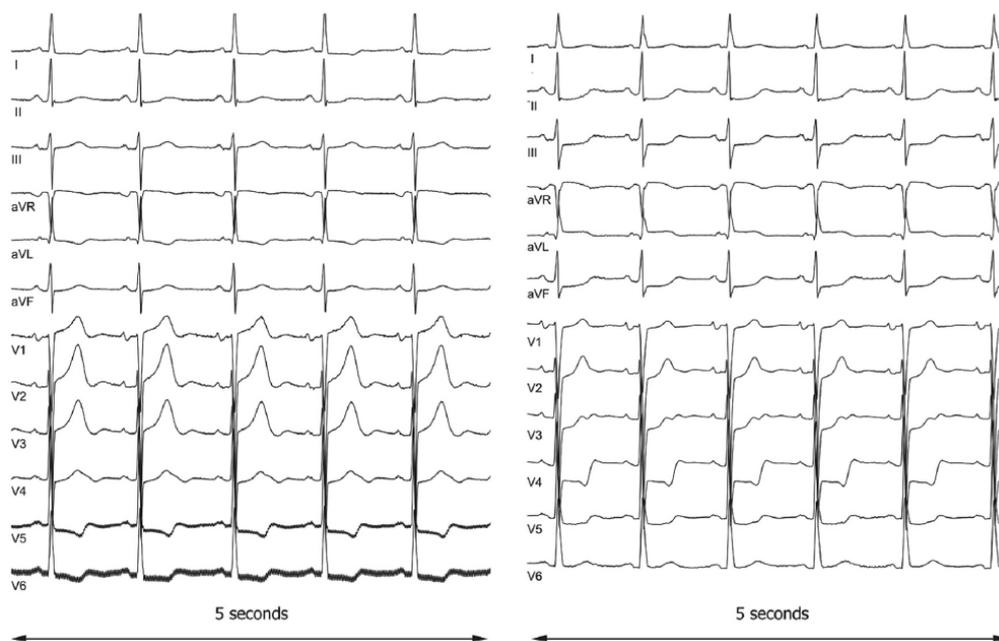


Fig. 8 – The ECG of a patient, suffering of left ventricular hypertrophy, with stable angina. Left panel: baseline ECG. Right panel: ischemic ECG after 3 min of balloon occlusion. From [47].

Obviously, in clinical practice, differential analysis of a preexisting ECG of the same patient requires that it has been made and stored in a database that is instantly accessible when the acute situation occurs. As such, the “age” of the historical ECG is important and should possibly kept limited. Serial comparison also faces the problem of variability in electrode placement between the historical and acute ECG, electrode misplacement [21]. In multiple cases, clear differences in

precordial P wave and QRS-complex orientation could be observed between the acute and reference ECG, suggesting that also J-point and T wave differences may be electrode-position related, hence, negatively affecting the final diagnosis [37].

Chapter 3 – Unsupervised Machine Learning

3.1 *Machine Learning and general workflow*

Machine learning (ML) broadly refers to the process of fitting predictive models to data or identifying informative groupings within data. Combining statistics and computer science, ML attempts to approximate or imitate the ability of humans to recognize patterns in an objective manner. ML is often confused with artificial intelligence (AI), neural networks, and deep learning [48]. ML is a subfield of AI, which defined as “a field of science and engineering concerned with the computational understanding of what is commonly called intelligent behavior and with the creation of artifacts that exhibit such behavior” [49].

Although the common perception for developing a ML application focuses on model training and evaluation, there are other steps essential to the successful development of generalizable and high performing ML models. The iterative process of an ML model development characterized by the following essential steps, also shown in *Fig. 9*:

- 1) *Data acquisition* is a fundamental step in developing any ML workflow. Missing key attributes or including confounding factors might lead to unintended consequences for the downstream analysis in model development.
- 2) *Exploratory Data Analysis* (EDA) refers to operations as calculating summary statistics and visualization of various features and their interactions. EDA could reveal characteristics of the data and set the directions for downstream analysis, such as data cleaning, feature selection and the choice of the most appropriate ML model.
- 3) *Data cleaning* refers to the processes of eliminating or amending the erroneous, corrupted, irrelevant, or missing values in a dataset, affecting the computational effort and the generalization capacity of ML model.
- 4) *Dimensionality reduction* refers to the methodologies used to transform the original high-dimensional data points into lower dimension ones while preserving the key properties of the original data points relevant to the task at hand. Dimensionality reduction methods are categorized as feature extraction and feature selection techniques. Feature extraction tries to combine the current features and provide a low-dimensional representation of the original data point. Methods based on Principal Component Analysis and Linear Discriminant Analysis are among the most widely used approaches. In contrast, feature selection refers to the methodology used for selecting a subset of the original features while preserving the key

properties of the original data points. The common families of feature selection methods are filter methods, wrapper methods, and embedded methods.

- 5) After conducting dimensionality reduction, the reduced or transformed set of features used to *train* ML models. When evaluating a ML model, available data are often partitioned into training, validation, and test sets. For small datasets, it is a common practice that 70% of the data points are used for training, 15% for validation, and 15% for testing. The training set used for model training, the validation set for hyperparameter tuning, and the test set for model evaluation. The purpose of learning the algorithm is to find patterns in the dataset and rate the data points according to those patterns.
- 6) Proper *evaluation* is essential for developing generalizable models. To provide an unbiased estimate of the generalization error, a test set locked away and is not used during model development, but only once in the model evaluation step. The learning algorithm tries to update the model parameters in a way that the performance measure improved [50].

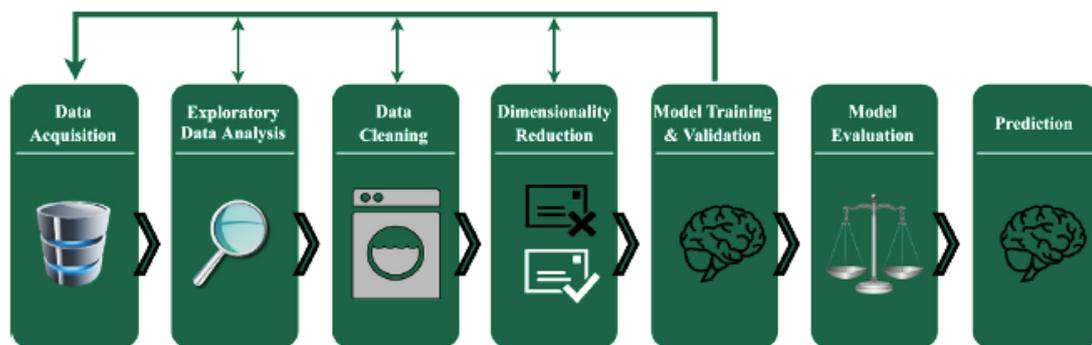


Fig. 9 – A general Machine Learning model development workflow. From [50].

Some potential applications of ML in healthcare include: predictive analytics, diagnosis and treatment, personalized medicine, clinical decision support, population health management. One of the challenges in building ML models is the data imbalance, also referred to as class imbalance. It is a crucial characteristic of many datasets, particularly in the healthcare domain, because of the rarity of a disease or phenotype [50]. This is current practice in healthcare applications, where the numbers of cases are often less than the number of controls [1]. Class imbalance can be measured as the difference between the size of the majority class and the size of the minority class. Models trained on imbalanced data may have a significant bias. Two common strategies for addressing class imbalance are oversampling and undersampling. In oversampling, observations from the minority classes selected using sampling with replacement to increase the number of samples from

the minority classes. Oversampling has also been criticized for changing the underlying distribution of the minority class, because a few observations in the minority class are used repeatedly to represent the data distribution for the minority class. In contrast, undersampling uses a bootstrapping approach, where a smaller number of observations from the majority class selected to decrease the class imbalance. The main shortcoming of the undersampling approach is that it discards a large number of samples from the majority class, thus it may lead to information loss. Given a set of labeled data points, ML models trained to provide a close estimate of the unknown mapping between inputs and outputs [50]. There are four main types of ML: supervised, unsupervised, semi-supervised and reinforcement learning. On the next section, the unsupervised learning techniques described.

3.2 Unsupervised Machine Learning

The labeling of the samples in a supervised ML is a very resource-intensive and time-consuming process, and usually, it can only be done manually, requiring certain expertise that is rare and expensive in domains as healthcare. So naturally it arises a need for methodologies that enable the training of such models with less or no labeled examples, goal achieved by unsupervised ML [51].

Unsupervised learning-based approaches play a crucial role in exploration of visualization-driven ECG data analysis. These approaches are useful for the detection of relevant trends, patterns, and outliers, which are not always amenable to expert labeling, as well as for identifying complex relationships between subjects and clinical conditions [7]. “Unsupervised” means that the machine or computer should learn patterns from the data without referring to any specific response. Indeed, in unsupervised learning, models trained to use unlabeled high-dimensional data and identify recurring patterns from the input data points, based on their intrinsic characteristics or similarity measures. The goal is to find associations and patterns among input dataset, reducing the chance of human error and bias, which could occur during manual labeling processes [48]. It starts from a similar framework as supervised learning, with instances (patients in this case) each characterized by a feature vector, where values are given for particular attributes. These data can be conveniently represented by a matrix, but instead of using this matrix to learn a model relating features to outcomes, it is exploited to find a group of patients who are similar to one another [52]. The key difference is that with supervised learning, a model learns to predict outputs based on the labeled dataset, meaning it already contains the examples of correct answers carefully mapped out

by human supervisors. Unsupervised learning, on the other hand, implies that a model tries to find any similarities, differences, patterns, and structure in data by itself, without human supervision [53]. Although domain knowledge can easily be embedded for supervised learning approaches (as the labeled data), incorporating such knowledge cannot easily be embedded into the unsupervised ML. These models cannot also be directly used for predictive modeling, because there is no outcome variable assigned to a single input data point [50]. Clustering, association mining and dimensionality reduction are among the most common uses of unsupervised ML. On the next section, the clustering methodology will be deepened.

3.3 Clustering

Clustering is labeled as an unsupervised ML technique, which allows a model to learn from unlabeled datasets and form homogeneous groups of data points based on (dis)similarity measures, called clusters. The three primary uses of data clustering are: underline the pattern and insight into the data, identify the degree of (dis)similarity between data points, data organization and summarization through cluster prototypes. The cluster analysis aims to organize a collection of data items into clusters, such that items within a cluster are more “similar” to each other than they are to items in the other clusters. The idea is to maximize intra-cluster similarity while minimizing inter-cluster similarity. A good clustering algorithm will create well separated clusters of data points with higher intra-cluster similarity but lower inter-cluster similarity [54].

The steps in a typical clustering process include feature selection, measure selection, data grouping and output evaluation. The notion of similarity can be expressed in hugely different ways, according to the purpose of the study, to domain-specific assumptions and to prior knowledge of the problem. Clustering is usually performed when no information is available concerning the membership of data items to predefined classes. Some criteria that provide significant distinctions between clustering methods and can help selecting appropriate method for one’s problem, are the following:

- *Objective of clustering*: some methods aim at finding a single partition of the collection of items into clusters. However, obtaining a hierarchy of clusters can provide more flexibility and other methods focus on this.

- *Nature of the data items*: most clustering methods developed for numerical data, but some can deal with categorical data or with both. Clustering numerical data is easier than clustering categorical data since there is no intrinsic similarity metric between category objects.
- *Nature of the available information*: many methods rely on rich representations of the data, which let one define prototypes, data distributions, multidimensional intervals, beside computing (dis)similarities. Other methods only require the evaluation of pairwise (dis)similarities between data items; while imposing less restrictions on the data, these methods usually have a higher computational complexity.
- *Nature of the clusters*: the degree of membership of a data item to a cluster is either in $[0, 1]$ if the clusters are fuzzy or in $\{0, 1\}$ if the clusters are crisp. For fuzzy clusters, data items can belong to some degrees to several clusters that do not have hierarchical relations with each other. This distinction between fuzzy and crisp can concern both the clustering mechanisms and their results. Crisp clusters can always be obtained from fuzzy clusters.
- *Clustering criterion*: clusters can be seen either as distant compact sets or as dense sets separated by low density regions. The compactness usually has strong implications on the shape of the clusters, so methods that focus on this aspect should be distinguished from methods that focus on the density [55].

The difference between classification and clustering is that clustering is conducted in an unsupervised manner, so no class labels are provided, and sometimes even the number of clusters is not known a-priory. In the case of unsupervised ML methods, the task is to uncover hidden relationships, structures, associations or hierarchies based on the data samples provided to the system. This kind of information gives us a better understanding of the data and the underlying processes generating it. The clusters can be constructed based on (dis)similarity or distance measures, so (dis)similar samples belong to the same cluster. The similarity measure can also be described as a distance measure, because the similarity of two samples can be interpreted as the distance between the two samples in the feature space. One approach for clustering is to use these measures, as degree of cohesiveness in an intra-cluster or separation in inter-cluster characteristics, and construct the regions of the feature space corresponding to the different clusters based on the computed distances between the training samples [51].

Defining X_i and X_j as two objects in cluster k , the most used distance metrics in accessing the (dis)similarity among the data objects, are:

- *Euclidean distance (ED)*: determines the root of square differences between the coordinates of a pair of objects and can be generalized to higher dimensions. Although ED is quite common in clustering, it has a drawback: if two data vectors have no attribute values in common, they may have a smaller distance than the other pair of data vectors containing the same attribute values. Another problem of ED as a family of the Minkowski metric is that the largest-scaled feature would dominate the others but can be solved with normalization. It obtained by Eq. (1).

$$Dist_{X,Y} = \sqrt{\sum_{k=1}^m (X_{ik} - X_{jk})^2} \quad (1)$$

- *Manhattan distance*: is determined as the total of the absolute differences between the two objects between any two points X and Y shown in Eq. (2). When this distance measure used in clustering algorithms, the shape of clusters is hyper-rectangular; moreover, Manhattan is sensitive to outliers, as ED.

$$Dist_{X,Y} = |X_{ik} - X_{jk}| \quad (2)$$

- *Chebyshev distance*: represents the maximum separation between any two points X and Y in a single dimension, according to Eq. (3).

$$Dist_{X,Y} = \max_k |X_{ik} - X_{jk}| \quad (3)$$

- *Minkowski distance*: is a generalized measure from which Manhattan, Euclidean and Chebyshev distances are particular cases, obtained by setting $p = 1$, $p = 2$ and $p = \infty$, respectively. The Minkowski distance performs well when the dataset clusters are isolated or compacted; otherwise, it has the same problem of ED with large-scale attributes. Minkowski distance metric represented by Eq. (4).

$$Dist_{X,Y} = (\sum_{k=1}^d |X_{ik} - X_{jk}|^p)^{1/p} \quad (4)$$

- *Cosine distance*: determines the similarity of two vectors A and B of n dimensions by calculating the cosine angle between them, as shown in Eq. (5). The cosine measure is invariant to rotation but is variant to linear transformations. It is also independent of vector length.

$$\dot{I} = \frac{\cos^{-1}(A \cdot B)}{\|A\| \|B\|} \quad (5)$$

- *Average distance*: to improve the clustering results, it is an updated version of ED, computed between two data points X and Y in n -dimension, defined as Eq. (6).

$$Dist_{X,Y} = \sqrt{\frac{1}{n} \sum_{k=1}^m (X_{ik} - X_{jk})^2} \quad (6)$$

- *Weighted Euclidean distance*: is a kind of ED that calculates the similarity between patterns when the weight (importance) of each feature specified as shown in Eq. (7). Calculating the weights is closely related to the dataset.

$$Dist_{weighted} = \sqrt{\sum_{k=1}^m w_i (X_{ik} - X_{jk})^2} \quad (7)$$

- *Mahalanobis distance*: in contrast to Euclidean and Manhattan distances which are independent of the linked dataset to which two data points belong, the Mahalanobis distance is a data-driven metric. By employing the squared Mahalanobis distance or performing a whitening adjustment to the data, Mahalanobis distance can reduce distortion brought on by linear correlation among features, useful in data classification and clustering. The mathematical form of Mahalanobis distance is defined in Eq. (8),

$$D_{mah} = \sqrt{(x - y)S^{-1}(x - y)^T} \quad (8)$$

where S is the covariance matrix of the dataset.

- *Pearson correlation*: this similarity metric determines how similar two gene expression patterns' shapes are. The Pearson correlation has a disadvantage of being sensitive to outliers. It defined by Eq. (9),

$$Pearson(x, y) = \frac{\sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^n (x_i - \mu_x)^2} \sqrt{\sum_{i=1}^n (y_i - \mu_y)^2}} \quad (9)$$

where μ_x and μ_y are the means for x and y, respectively.

- *Jaccard distance*: Jaccard distance, a traditional similarity measure on sets, useful in information retrieval, data mining, ML. The Jaccard distance between two finite sets A and B shown in Eq. (10) [54].

$$J(A, B) = \left| \frac{A \cap B}{A \cup B} \right| \quad (10)$$

- *Braycurtis distance*: this distance measure is bounded between 0 (when the sample units are identical) and 1 (when the sample units are completely different), and is semi-metric, since it does not satisfy the triangle inequality axiom. Supposing that the counts are denoted by n and that their sample (row) totals are n_+ , the computation involves summing the absolute differences between the counts and dividing this by the sum of the abundances in the two samples i and j, as shown in Eq. (11) [56].

$$b_{ij} = \frac{\sum_{m=1}^M |n_{im} - n_{jm}|}{n_{i+} + n_{j+}} \quad (11)$$

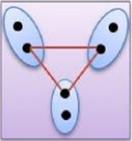
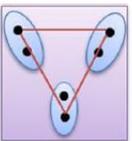
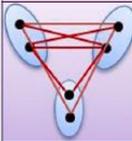
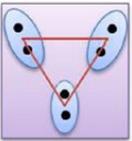
To obtain optimal clustering results, cluster analysis faces different challenges at distinct phases. No one technique can be expected to work effectively for all types of data since clustering techniques are subjective in nature. There are many challenges involved with regards to selecting parametric characteristics when addressing the clustering problems. These parametric characteristics are broadly classified into two categories: choice of datasets and selection of computational methods [54].

In clustering, every dataset instance will receive a label indicating its participation in a cluster. The base of clustering algorithms is often the same, although the methods used to calculate distance and (dis)similarity as well as how labels are chosen frequently vary. There are three categories of clustering algorithms: partitional, density-based and hierarchical clustering.

- Partitional clustering, also known as iterative relocation algorithm, concurrently divides the dataset into homogeneous groups without creating a hierarchical structure. Partitional clustering attempts to optimize a specified criteria function, reflecting the “agreement” between the data and the partition. The objective function optimized to iteratively partition the entire dataset into a predetermined k number of groups. Algorithm’s key advantages are that it is easy to build and that it converges quickly [54]. Partitional clustering divided into two categories: hard/crisp clustering and fuzzy/soft clustering. Hard clustering occurs when data points belong to one cluster, whereas data points belonging to one or more clusters referred to as soft clusters [57].
- Density-based clustering is a nonparametric approach that uses the idea of density to discover clusters of various forms, sizes, and densities. It is a traditional and popular clustering method to extract hidden patterns from datasets by separate high- and low-density regions based on the neighborhood information. Usually, data objects found in low-density regions considered as noise or outliers. There are typically two key phases in density-based approaches. In the first step, based on the local neighborhood information, it calculated the density of each data point. After that, similar data points in denser regions are identified and combined with them to build clusters. A prominent example of this category is Density-Based Spatial Clustering (DBSCAN), which successfully manages outliers and can identify clusters of any shape. However, its implementation relies on two user inputs: *epsilon*, representing the maximum radius of a point in a dataset to measure the density; *minpts*, indicating the minimum number of points needed inside a circle of radius epsilon distance for that data point to be categorized as a core point. Due to the significant parameter sensitivity of clustering algorithms, meticulous manual parameter adjustment, the application of DBSCAN is so limited [54].

- Hierarchical clustering aims to obtain a hierarchy of clusters, called dendrogram, that shows how the clusters are related to each other. These methods proceed either by iteratively merging small clusters into larger ones (agglomerative algorithms, by far the most common) or by splitting large clusters (divisive algorithms) according to specific criteria. Different algorithms may be implemented depending on the way similarity is measured between clusters, such as single linkage, average linkage, complete linkage, and centroid linkage methods, summarized in *Table IV*. Another popular linkage is *Ward* which examines multivariate Euclidean space for clusters [55].

Table IV – Linkage criteria used in the hierarchical clustering. Modified from [54].

<i>Single linkage</i>	<i>Complete linkage</i>	<i>Average linkage</i>	<i>Centroid linkage</i>
			
This technique joins two clusters with the smallest member distance.	This technique joins two clusters with the longest member distance. This linkage is less sensitive to noise and produces compact clusters.	This technique joins two clusters with the smallest average member distance. It is commonly referred to as the minimum variance linkage, so preferable than single and complete links.	This technique joins two clusters with the smallest centroid member distance. When dealing with clusters of various sizes, this linking approach typically outperforms others, since it is more resilient to outliers.

The clustering methods involved in this work for comparative analysis are described in-depth in the following subsections.

3.1.1. K-Means

K-means clustering aims to partition the data points of a dataset into k clusters, number defined a-priori. This partitioning accomplished by first initializing the center of each cluster based on an initialization strategy. One strategy is to randomly choose n observations from the dataset and select these data points as the initial

centroids of the k clusters. In an iterative manner, each time all data points in the dataset assigned to one of the k clusters based on their proximity to the cluster centroids, according to the similarity measure chosen. Next, each cluster centroid updated to be the average of all data points in its corresponding cluster. This process continues until there is no change in the cluster centroids or a stopping criterion is satisfied. The stopping criteria can be that no further change in the classification of training samples happened after the last update of the cluster centroids, or the distance between the centroids before and after the update is smaller than a specified value. K-means is a scalable approach with guaranteed convergence. However, the converged solution might be non-optimal. K-means can easily adapt to the changes in data distribution by introducing the new data points and updating its centroids. On the negative side, K-means is heavily constrained by the proper selection of preliminary parameters, like the number of clusters or the initial location for the centroids. It is computationally intensive to calculate the distance of each sample and each centroid, so for a large number of samples, the basic algorithm has to be altered. Note that K-means requires the features to be homogeneous and numerical; otherwise, the distance between data points might be driven by the features with large absolute values [50] [51]. K-means also starts with a random choice of cluster centers and therefore it may yield different clustering results on different runs of the algorithm. Thus, the results may not be repeatable and lack consistency [53].

3.1.2. K-Medoids

Partition Around Medoids (PAM) is developed by Kaufman and Rousseeuw in 1987 and it is based on classical partitioning process of clustering. K-Medoid Clustering algorithm is like the K-means clustering algorithm, but the difference lies in the center point, the medoid, instead of centroid. K-medoids uses an actual object to find the most central object within the cluster and assign the nearest object to the medoids to create a cluster, instead of using the mean value of an object in the cluster as a reference point. The algorithm begins with arbitrary selection of the K objects as medoid points out of n data points ($n > K$), then swaps the medoid object with non-medoid, thereby improving the quality of cluster. After selection of the K medoid points, associate each data object in the given data set to most similar medoid and then randomly select non-medoid object O . At this point, it needed to

compute the total cost S of swapping initial medoid object O . If $S > 0$, swap initial medoid with the new one. Repeat steps until there is no change in the medoid. K-Medoids can find the most centrally located point in the given dataset, since it is more robust to outliers and noises, because of the Manhattan distance as metric. Further, it employs best practices when scaling large datasets and it is a fast-clustering method of categorical data. However, K-Medoids is more costly than K-Means method, because of its time complexity, thus it does not scale well for large datasets [54] [58].

3.1.3. Spectral Clustering

Spectral clustering used for graph partitioning by analyzing graphs with methods of linear algebra. Spectral clustering unravels the structural properties of a graph using information conveyed by the spectral decomposition (eigen decomposition) of an associated matrix. The elements of this matrix code the underlying similarities among the nodes (data points) of the graph. The training data can be represented as a similarity graph, which is an undirected graph, with the training samples as the vertexes and the edges associated with a weight of the similarity between the two vertexes they connect. The similarity function needs to be symmetric and non-negative [51]. When constructing similarity graphs, the goal is to model the local neighborhood relationships between the data points. Indeed, the two mathematical objects used by spectral clustering are similarity graphs and graph Laplacians, with the latter computed from the former. There exists a whole field dedicated to the study of those matrices, called spectral graph theory [59]. The central idea behind Spectral Clustering lies in the matrix, which summarizes the relationships between data points within the graph. There are two types of matrices: an abnormalized Laplacian and a normalized Laplacian. The first one defined as in *Eq. (12)*,

$$L = D - W \quad (12)$$

where W is the graph's adjacency matrix, a square matrix that contains the similarity between each pair of nodes, and D is the graph's degree matrix, a diagonal matrix that contains the number of edges associated to each node. An overview of many of its properties can be found in [60][61]. There are two matrices which called

normalized graph Laplacians in the literature, both closely related to each other and defined as *Eq.s (13) (14)*,

$$L_{sym} = D^{-1/2}L D^{-1/2} = I - D^{-1/2}W D^{-1/2} \quad (13)$$

$$L_{rw} = D^{-1}L = I - D^{-1}W \quad (14)$$

where L_{sym} as it is a symmetric matrix and the second one by L_{rw} as it is closely related to a random walk. The standard reference for normalized graph Laplacians is [59].

Firstly, it needed to construct a similarity graph, following ϵ -neighborhood graph, k-nearest neighbor graphs or fully connected one, described in *Section 2* of [62]. Next, an unnormalized or normalized Laplacian computed according to *Eq. (12) (13) (14)*, for capturing characteristic aspects of the underlying graph. To effectively transform data into a novel space that allows for enhanced separation between clusters, it exploited the eigenvalues decomposition of the Laplacian matrix chosen. Finally, it is necessary to apply a clustering algorithm, as K-means, to the embedding of the nodes defined by the eigenvectors. Spectral Clustering is extremely useful when the structure of the individual clusters is highly non-convex, or more generally when a measure of the center and spread of the cluster is not a suitable description of the complete cluster. It is simple to implement, can be solved efficiently by standard linear algebra software, and very often outperforms traditional clustering algorithms. Moreover, spectral clustering can be implemented efficiently even for large data sets, if the similarity graph is sparse. It is relatively robust to noisy data, too. Once the similarity graph is chosen, there is the need to just solve a linear problem, and there are no issues of getting stuck in local minima or restarting the algorithm for several times with different initializations. However, choosing a good similarity graph is not trivial and spectral clustering can be quite unstable under different choices of the parameters for the neighborhood graphs, as the number of cluster and the affinity measure. Additionally, it can be computationally intensive, especially when dealing with large datasets due to eigenvalue decomposition step [62].

3.1.4. Agglomerative Clustering

The agglomerative clustering strategy approach is known as “bottom to up”, directing “the leaves” to “the root” of a cluster tree. In this approach, every single data point is firstly assigned to a separate cluster and form a singleton set. Then, in an iterative process, these clusters start merging until there is 1 cluster containing all n data points. In each iteration, the two most similar (nearest) clusters selected and merged. The similarity between clusters quantified by a linkage criterion, among those listed in *Table IV*, and a distance metric, defined in the previous section. The metric used for quantifying the distance between data points, and the linkage method for measuring the similarity between clusters [50].

It builds a hierarchical structure of clusters, which can be represented as a dendrogram. The leaves of the dendrogram structure are the samples themselves (each belonging to its own class), and the root of the structure is the cluster that includes all the samples. Thus, cutting the dendrogram at various levels of hierarchy results in a different number of clusters. Users have the option to choose the relevant groupings of characteristics from the dendrogram based on various criteria, and the relationships between the groups are preserved. Unlike the K-means algorithm, hierarchical clustering does not propose disjoint clusters and does not require the a-priory declaration of the number of clusters, however doing so can serve as a stopping condition, resulting in faster computation for the proposed clusters [51]. Moreover, it is less sensitive to noise and outliers in the datasets. Along with these main advantages, hierarchical clustering has limitations associated with it. In terms of time and memory, it is computationally expensive, especially in larger datasets. In addition to this, it is not easy to decide the dendrogram level in this approach, where clusters rely on the distance metric used [54].

3.1.5. CLARA

Clustering LARge Applications (CLARA) has been developed by Kaufman and Rousseuw in 1990. This partitioning algorithm has come into effect to solve the problem of PAM. It is an extension of K-Medoids clustering algorithm, which uses the sampling approach to handle large datasets. In CLARA, a sample of the data set is chosen from the entire data and then PAM is used to select arbitrarily medoids

out of the sample. The concept is that if the sample is chosen in a fairly random manner, then its medoid will correctly represent the whole data set [63].

First, it is necessary to draw a sample of n objects randomly from the entire data set and call PAM to find k medoid of the sample. Then, for each sample the specific K medoid which is similar to the given object is determined. Calculate the average dissimilarity of the clustering thus obtained and if the value is less than the present minimum, it maintained and retained the K -Medoid found in the previous step as best of medoid. The most important advantage of CLARA is that runs on multiple samples and gives the best clusters out of the given set of samples. The efficient performance of CLARA depends upon the size of the dataset and on the a-priori knowledge of the number of clusters. It reported by Lucasius, Dane, and Kateman in [64] that the performance of CLARA drops rapidly below an acceptable level with an increasing number of clusters. Additionally, biased sample data may result in misleading and poor clustering of whole datasets [58].

3.1.6. Fuzzy C-Means

The Fuzzy C-Means (FCM) clustering algorithm is a popular approach that clusters data points when it belongs to more than one cluster. FCM is an implementation of a fuzzy clustering method which uses K-Means principles to divide a dataset into clusters. It is suitable for pattern recognition when clusters overlap: a sample has a degree of membership in each cluster, which is a continuous value rather than a binary. Where N is the number of samples, C is the number of clusters, x_i is the i^{th} sample, $i \in \{1, 2, \dots, N\}$, c_j is the j^{th} cluster centroid, $j \in \{1, 2, \dots, C\}$, μ_{ij} is the degree of membership of x_i in cluster j , $\| \cdot \|$ is any norm for measuring distance (like Euclidean norm) and m is a coefficient to control fuzziness $1 \leq m \leq \infty$. The FCM clustering is produced by minimizing the objective function J_m , as given in Eq. (15), with an iteration process, during which the degrees of membership for each sample and each cluster are updated.

$$J_m = \sum_{i=1}^N \sum_{j=1}^C \mu_{ij}^m \|x_i - c_j\|^2 \quad (15)$$

The degrees of membership, ranged from 0 to 1, reflect its membership value to a cluster rather than being assigned to a single cluster exclusively, can be determined according to *Eq. (16)*.

$$\mu_{ij} = \frac{1}{\sum_{k=1}^C \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \quad (16)$$

The preliminary steps before the iterative algorithm are to define C , the coefficient for fuzziness (which is usually set to 2) and to assign an initial degree of membership for all training samples for each cluster. This is usually done by filling a matrix U of size $N \times C$ with random values for μ_{ij} . After these preliminary steps, the cluster centroids computed with *Eq. (17)*, based on the training samples and the given U matrix containing the degree of membership of each training sample in each cluster.

$$c_j = \frac{\sum_{i=1}^N \mu_{ij}^m \cdot x_i}{\sum_{i=1}^N \mu_{ij}^m} \quad (17)$$

Then, the elements of the matrix U modified according to *Eq. (16)*. These two steps are repeated until the stopping condition, formulated exactly like in the k-mean algorithm, is met [51] [54].

The strength of FCM is in how it allows clusters to assign flexibility, where it is more practical to provide the probability of belonging to a cluster. However, this algorithm has some weaknesses relating to high complexity in specifying the number of clusters in advance, unable to handle high-dimensional datasets, cluster validity, and a priori knowledge of cluster shape and density [54].

3.1.7. Gaussian Mixture Model

Gaussian Mixture Model (GMM) is one of the different variations of density-based algorithm class, among those in the literature [54]. Many real-life datasets comprise different clusters, thus fitting a probabilistic model for such data cannot be evaluated by one distribution, like a Gaussian. Intuitively, one Gaussian distribution fits all data points in a shape, as a circle or ellipsoid, and it assigns a higher probability to the center of the shape which is not applicable for multi-cluster datasets with multi-centers. Accordingly, multi-Gaussian (a sequence or mixture) can model multi-class datasets to meet the number of clusters [65]. The GMM is an

extension of a single Gaussian probability density function, which uses multiple Gaussian probability density functions (normal distribution curves) to quantify the distribution of variables accurately. This decomposes the variable distribution into several Gaussian probabilities of the statistical model of densities function distribution. The model adjusts the means, coefficients, and covariance through enough Gaussian distributions to approximate any continuous function of density closely. The GMM can effectively capture the internal correlation structures within datasets, as it is a flexible model for a wide range of distribution probabilities. In addition, it has high accuracy and real-time implementation. The drawbacks of the Gaussian mixture model relate to it being computationally expensive with large distributions or with few observed data points in datasets. Further, it can be difficult to estimate the number of clusters, which require large datasets [54].

3.1.8. BIRCH

Zhang developed BIRCH, a clustering algorithm suitable for large datasets [66]. The proposed method is relied on clustering features (CF) and CF tree. CF is a tuple of (N, LS, SS) , where N is the number of data points in the cluster, LS is the linear sum of N data points and SS is the square sum of N data points. By building the CT tree, in which one node represents a subcluster, BIRCH realizes the clustering result. The CF tree expands dynamically when a new data point added to it. BIRCH is local (as opposed to global) in that each clustering decision made without scanning all data points or all currently existing clusters. It uses measurements that reflect the natural closeness of points, and at the same time, can be incrementally maintained during the clustering process. It is an incremental method that does not require the whole dataset in advance and only scans the dataset once. The capacity to manage big datasets and resilience to outliers are the two key motivations. However, the effectiveness of BIRCH depends on proper parameter setting. The BIRCH clustering method can provide high-quality clustering at a less computational cost. The algorithm includes several advantages such as: (A) using an incremental clustering technique, creates a CF tree from a single scan of the dataset; (B) able to efficiently manage noise; (iii) memory-efficient method, because it saves only a limited number of abstracted data points rather than the entire dataset [54]. BIRCH makes full use of available memory to derive the finest

possible subclusters (to ensure accuracy) while minimizing I/O costs (to ensure efficiency). The clustering and reducing process is organized and characterized by the use of an in-memory, height-balanced and highly occupied tree structure. By evaluating time/space efficiency, data input order sensitivity, and clustering quality, and comparing with other existing algorithms through experiments, BIRCH is the best available clustering method for exceptionally large databases. BIRCH's architecture also offers opportunities for parallelism and for interactive or dynamic performance tuning based on knowledge about the dataset [66].

3.1.9. Self-Organizing Map

Around 1981–82, Kohonen [67] introduced a new nonlinearly projecting mapping, called the self-organizing map (SOM), which otherwise resembles the vector quantization (VQ), but in which, additionally, the models become spatially, globally ordered. SOM is a type of artificial neural network that trained using unsupervised learning to produce a low-dimensional, discretized representation of the input space of the training samples, called a map, consisting of rows and columns, if two-dimensional [13]. The SOM algorithm constructs the models such that: “more similar models will be associated with nodes that are closer in the grid, whereas less similar models will be situated gradually farther away in the grid” [67]. The formation of an SOM involves three characteristic processes: competition among the output nodes (neurons), cooperation due to nearby locations in the output space representing inputs with similar properties, adaptation of the weight vectors of the winner with its neighboring units [68].

The nodes in the input layer denote the attributes (features) or, more generally, the variables contained in the input data. These vectors are associated with one of the different clusters. The input layer of source nodes is directly connected to the output layer of computation nodes without any hidden layer. In the output layer of the network, also known as “Kohonen layer” or “SOM layer”, there is an output neuron for every cluster, representing a low dimensional visualization of the data. The number of nodes in the output layer denotes the maximum number of clusters and influences the accuracy and generalization capability of SOM. So, the output layer has k number of neurons, a layer that is parameterized by a set of weights, which are trained using an online algorithm to fit a set of prototypes c (similar to centroids

of clusters) to the data. The neuron with the highest activation decides the cluster a sample belongs to. The weight vector with the minimum distance from the sample can also be found by finding the maximum of the scalar products of the input vector and the weight vectors if the weight vectors are all normalized [51].

The formation of SOM starts first by initializing the weight vectors $w_i = (w_{i1} w_{i2} \dots w_{in})'$ where $i = (1, 2, \dots, n)$ and denotes the number of output nodes in the network. Weights are links that connect the input nodes to the output nodes and updated through the learning process. For a given dataset X , the SOM selects a random sample x_i at time, the distances between the sample and all prototypes computed at each iteration t by applying the minimum distance criterion, as in Eq. (18),

$$c(t) = \min \|x(t) - w_i(t)\| \quad (18)$$

where c_r is the best matching node, corresponding to the location of the center neuron that maximized to be indicated as a winner neuron. Subsequently, the weights updated, taking into consideration two parameters: learning rate and neighborhood size. The learning rate controls the rate of change of the weight vectors and, as in all neural networks, it takes values between 0 and 1. In SOMs, the learning rate gradually decreases as a function of the iteration step index t . Neighborhood information, important for preserving the topological properties of the original data in the output space, can be expressed as a function, chosen to update the weight vectors of the respective nodes. The neighbors for the i^{th} neuron identified such that in Eq. (19),

$$nb = \exp\left(\frac{\|c_i - c_r\|^2}{\omega^2}\right) \quad (19)$$

where ω is the variance of nb . The output nodes compete among themselves to become activated. Only the node whose weight vector is most similar to the input vector will be activated and declared as the winner. To find this best matching node, the distances between an input data x and all the weight vectors w_i of the SOM computed using different measurement methods, listed in the previous section. Of all the methods, as Kohonen points out, Euclidean distance is more favorable in visual representations like those in SOMs, because a more isotropic display obtained using it. The spaces between the neighbors are arbitrary and not important for the system. The focus is on the adjacency for two clusters that shows the similarity [65]. The stranded SOM neglects the loss function form the model

optimization; conversely, it is considered in the probabilistic SOM, that has been introduced in view of Gaussian-full mixture components [69].

SOM can produce simpler low-dimensional representations of complex high-dimensional data at the expense of information loss during the projection. Moreover, the outputs of an SOM can vary highly depending on the initial setting of parameters such as the number of output nodes, the learning rate, and the update speed of the neighborhood. In comparison to the multivariate techniques, the nonparametric SOM procedures have many advantages. First, they do not make assumptions regarding the distributions of variables and nor do they require independence among variables. Second, they are easier to implement and can solve nonlinear problems of very high complexity. Last, they more effectively cope with noisy and missing data, small dimensionality and samples of unlimited size [68].

3.1.10. Mean Shift

The mean shift algorithm is a powerful general non-parametric mode finding procedure. It is a density-based clustering technique, where each point is assigned to the nearest cluster center. More specifically, it is a centroid-based algorithm, which works by updating candidates for centroids to be the mean of the points within a given region. These candidates are then filtered in a post-processing stage to eliminate near-duplicates to form the final set of centroids. The main idea of the mean shift is to treat the points in D-dimensional feature space as an empirical probability density function where dense regions correspond to the local maxima of the underlying distribution. The mean shift procedure consists of two steps: firstly, construction of probability density in some feature space, secondly the mapping of each point to the maximum (mode) of the density which is closest to it. Each data point is shifted to the weighted average of the data set. The mean shift algorithm is an iterative process which computes the mean shift value for the current position and then moves the point to its mean shift value. Gradient ascent is performed in the feature space on the local density estimation until convergence. After the procedure, stationary points correspond to the modes of the distribution, and the same stationary points are considered members of the same cluster. In contrast to the well-known K-means clustering approach, mean shift does not need assumptions on the number of clusters and the shape of the distribution, but its performance

relies on the selection of scale parameters. Bandwidth is the only parameter to tune, so for the one-dimensional case this is a relatively simple procedure, but in a multidimensional case, it can be difficult.

Among the strengths of the mean shift algorithms we have that it does not assume any prior shape on data clusters, and it can handle arbitrary feature spaces. The algorithm is not sensitive to outliers, and convergence guaranteed. Weaknesses of the mean shift algorithm include a need to use adaptive window size because improper window size can lead modes to be merged, which results in bad clusters (segments). The inappropriate window size can cause modes to be merged or generate additional shallow modes. A limitation of the standard mean shift procedure is that the value of the bandwidth parameter is unspecified. Additionally, mean shift might not work well in higher dimensions. Computational complexity depends on the number of iterations and the number of data points, while the convergence speed and quality on the kernel type. The most computationally expensive component of the mean shift procedure corresponds to identifying the neighbors of a point in a space, becoming cumbersome for high-dimensional feature spaces [70].

K-means, K-medoids, Spectral, CLARA, FCM and BIRCH clustering techniques classified as partitional methods, while GMM and Mean Shift as density-based category. Differently to Agglomerative clustering and SOM, which are the most used algorithms belonging to hierarchical clustering and artificial neural network, respectively.

Chapter 4 – Materials & Method

4.1 Database

Data were collected during the SUBTRACT study, described in [37], jointly performed by the Amsterdam Medical Center (AMC) and the Leiden University Medical Center (LUMC), a retrospective observational study that aimed to evaluate the diagnostic value of serial electrocardiography for the detection of acute myocardial ischemia in the pre-hospital phase. The demographic and anthropomorphic characteristics of the study group involved as well as the medical history can be also found in [37]. A couple of 10-s 12-lead ECGs were acquired from patients that performed an urgent ECG exam in an ambulance for ruling in/out myocardial ischemia. Each couple is composed of one ambulance ECG (AECG) made in an acute situation and one electively made prior ECG that functions as a reference ECG (RECG). The AECG corresponded to the performed ECG during the urgent ECG exams in ambulance and was recorded with LIFEPAK 12 (Physio-Control) using the Mason-Likar electrode configuration (Mason and Likar, described in-depth in [71]). The RECGs were an early performed ECG of the same patient in non-acute conditions and were, depending on the hospital in which they were made, recorded by various electrocardiographs (GE, Schiller, Mortara, Siemens/Dräger) using the standard electrode configuration of the diagnostic resting ECG [1]. The previously recorded RECGs had a median “age” (time difference between the recording of the AECG and of the RECG) of 12 months (minimum: 0, Q1: 4, Q3: 34, maximum 332 months). In patients in whom multiple AECGs had been recorded, the AECG with the largest magnitude of the ventricular gradient (VG) difference vector with respect to the RECG was selected for further analysis. This choice motivated by the notion that VG differences reflect changes throughout the QRST-complex. To retrospectively assess the presence or absence of myocardial ischemia at the time of recording of the AECG, it was defined and applied a myocardial ischemia classification algorithm. The algorithm is based on interpretation of the clinical in-hospital data, with the purpose of retrospectively constructing the prehospital scenario. For this purpose, the algorithm uses a 5-point scale, ranging from presumed ischemic, probably ischemic, uncertain, probably nonischemic to presumed nonischemic, described in-depth in [37]. A total of 1425 patients were included, subdivided in 736 males and 689 females (67 ± 14 years). Considering the exclusion criteria reported in [37], the resulting study group comprised 194 (14%) cases and 1035 (73%) controls, as summarized in the following *Table V*. The differences between the AECG and RECGs of each patient were expressed as difference descriptors, each of which was obtained by subtracting LEADS RECG output

variables from LEADS AECG output variables [1]. Features, listed in *Table VI*, were grouped into two features set: (1) direct measurements, from 1 to 18, and (2) serial features, from 19 to 47.

Table V – Ischemia categories, numbers of patients and case-control assignment in the study group. From [37].

Ischemia Category	Number (%) of patients	Target
<i>Presumably ischemic</i>	166 (12%)	Cases
<i>Probably ischemic</i>	28 (2%)	Cases
<i>Uncertain</i>	196 (13%)	Excluded
<i>Probably non-ischemic</i>	66 (5%)	Controls
<i>Presumably non-ischemic</i>	969 (68%)	Controls
	1425 (100%)	

Table VI – List of features computed from the Ambulance ECG (AECG) and Reference ECG (RECG). Adi: adimensional. From [37].

#	Direct AECG Measurements	Description	#	Serial AECG-RECG features	Description
1	QRS (ms)	QRS duration	19	QRSD (ms)	Difference between the QRS durations
			20	QRSaD (ms)	Absolute value of the difference between the QRS durations
2	MQRS (μV)	Magnitude of the maximal QRS vectors	21	MQRSD (μV)	Difference between the magnitudes of the maximal QRS vectors
			22	MQRSaD (μV)	Absolute value of the difference between the magnitudes of the maximal QRS vectors
3	IQRS (mV·ms)	Magnitude of the QRS-integral vector	23	IQRSD (mV·ms)	Difference between the magnitudes of the QRS-integral vectors
			24	IQRSaD (mV·ms)	Absolute value of the difference between the magnitudes of the QRS-integral vectors
4	QRSC (%)	QRS complexity	25	QRSCD (%)	Difference between the QRS complexities
			26	QRSCaD (%)	Absolute value of the differences between the QRS complexities
5	J (μV)	Magnitude of the J-point vector	27	JD (μV)	Absolute value of the difference between the magnitudes of the J-point vectors
6	SJ8 (μV)	Summed absolute J-point amplitudes considering the 8 independent leads	28	SDJ8 (μV)	Summed absolute values of the relative J-point shifts considering the 8 independent leads

7	SJ12 (μV)	Summed absolute J-point amplitude considering all 12 leads	29	SDJ12 (μV)	Summed absolute values of the relative J-point shifts considering the 12 standard leads
8	MT (μV)	Magnitude of the maximal T vector	30	MTD (μV)	Difference between the magnitudes of the maximal T vectors
			31	MTaD (μV)	Absolute value of the difference between the magnitudes of the maximal T vectors
9	IT ($mV \cdot ms$)	Magnitude of the T-integral vector	32	ITD ($mV \cdot ms$)	Difference between the magnitudes of the T-integral vectors
			33	ITaD ($mV \cdot ms$)	Absolute value of the difference between the magnitudes of the T-integral vectors
10	TC (%)	T-wave complexity	34	TCD (%)	Difference between the T-wave complexities
			35	TCaD (%)	Absolute value of the difference between the T-wave complexities
11	TS (%)	T-wave symmetry	36	TSD (%)	Difference between the T-wave symmetries
			37	TSaD (%)	Absolute value of the difference between T-wave symmetries
12	NPT (<i>adi</i>)	Number of leads with positive T waves	38	NPTD (<i>adi</i>)	Difference between the number of leads with positive T waves
			39	NTPC (<i>adi</i>)	Number of leads that present a T-wave polarity change
13	QT (<i>ms</i>)	QT interval	40	QTD (<i>ms</i>)	Difference between the QT intervals
			41	QTaD (<i>ms</i>)	Absolute value of the difference between the QT intervals
14	VG (μV)	Magnitude of the ventricular gradient	42	VGD (μV)	Difference between the magnitudes of the ventricular gradients
15	SA ($^{\circ}$)	QRS-T spatial angle	43	SAD ($^{\circ}$)	Difference between the QRS-T spatial angles
			44	SAaD ($^{\circ}$)	Absolute value of the difference between the QRS-T spatial angles
16	HR (<i>bpm</i>)	Heart rate	45	HRD (<i>bpm</i>)	Difference between the heart rates
			46	HRaD (<i>bpm</i>)	The absolute value of the difference between the heart rates
17	Age	Age	47	AgeD	Age of the RECG
18	Sex	Sex			

Another database was created from the same SUBTRACT study [37], in which the target variable indicates the presence or absence of ACS, while maintaining the same 46 features. The resulting databases are composed as follows: target (where 0 and 1 indicate controls and cases, respectively), 18 direct ECG measurements, collected during the acute phase, in which sex and age are included, and finally 28 serial ECG features.

4.2 Implementation

For this work, it was used Python as the programming language and executed the implementation in a Google Colab environment. Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing as well binding, makes it extremely attractive for rapid application development, in data science and ML. On the other hand, Google Colab is a free cloud-based platform that enables users to write and execute Python code collaboratively in a Jupyter Notebook environment, offering free Graphics Processing Unit (GPU) resources access, no setup requirements, and easy sharing for ML and data science tasks. The workflow for clustering the databases is shown in Fig. 10.

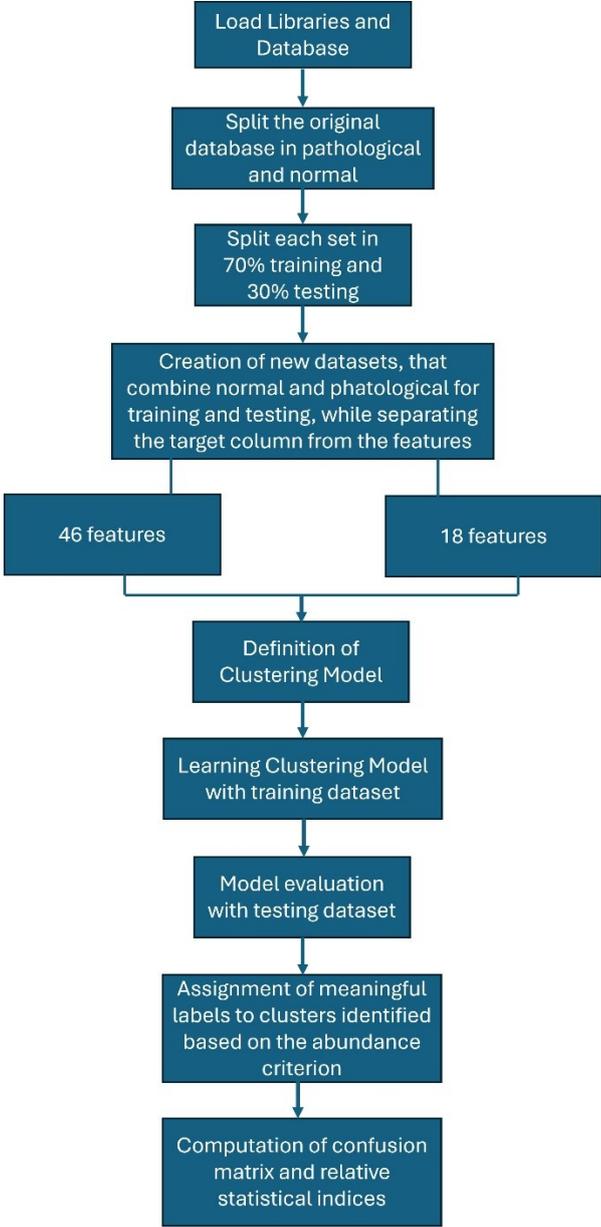


Fig. 10 – The followed workflow for clustering the myocardial ischemia and ACS databases.

Since Python supports modules and packages, which encourages program modularity and code reuse, it was firstly needed to import, and eventually install by using the package manager *pip*, libraries, that are collections of pre-written code that extend the functionality of the Python programming language. The common modules among the algorithms are: *pandas* for data manipulation and analysis; *numpy* for numerical operations, collections for exploiting the function counter in the labels' assignment phase; *sklearn* for exploiting its functions when implementing most of clustering algorithms and computing the confusion matrix for the performance evaluation. The exceptions are FCM, K-medoids and CLARA from *fuzzy_c_means* and *scikit_learn_extra* libraries, respectively. The package *sklearn* was also used for computing the distance and affinity matrixes according to the appropriate distance metric chosen, in Agglomerative and Spectral Clustering respectively, due to the absence of *predict* method, required for the testing evaluation. Next, the two databases, one containing data of subjects diagnosed as ACS patients and the other with myocardial ischemia, described in-depth in the previous section, were sourced from Excel files. These files were then imported into a Python script by connecting to a Google Drive account and read using the *pandas* library function. Then, there were extracted the rows corresponding to normal and pathological from the column of the target, 0 and 1 respectively. This phase is needed to divide then each set into 70% training and 30% testing data to firstly instruct and consequently evaluate how well the clustering model performs on unseen and unknown data, but to models comparison too. Once split each set randomly, there were created two new variables, one containing only normal and pathological belonging to the training partition and the other with those concerning the testing one, by exploiting the function *concat* of *pandas* library, where it was used the option '*ignore_index=True*' to reset the index in the resulting new dataframe. The methods were evaluated in two steps: first by considering all 46 features, and then by focusing on just the first 18. This approach aims to assess model performance in relation to the input data type, particularly the significance of serial features in disease detection. Subsequently, the clustering algorithm and its associated settings, detailed in *Table VII*, were adjusted according to the errors, so-called parameter tuning, considering both ACS and ischemia databases and utilizing 46 and 18 features. Particularly in libraries like *scikit-learn*, the *random_state* parameter, settled generally to a specific integer, is often required in functions that involve some form of randomness, such as data splitting, initializing models, or random sampling. The purpose of this parameter is to ensure reproducibility of results, which is particularly useful for debugging and sharing results, as it guarantees that random processes produce the same output each time the code runs. A common parameter in clustering algorithms is *K*, which is set to 2 to indicate that the goal is to identify two final clusters: one for pathological cases and one for normal cases.

Table VII – Clustering Methods: libraries, function, parameters setting.

<i>Clustering Method</i>	<i>Python Libraries</i>	<i>Function</i>	<i>Parameters</i>	<i>Random_state =0</i>
<i>K-means</i>	sklearn.cluster	KMeans	K	Yes
<i>K-medoid</i>	Scikit-learn-extra	KMedoids	K, metric = 'correlation', method = 'PAM'	Yes
<i>Spectral Clustering</i>	sklearn.cluster	SpectralClustering	K, affinity = 'cosine', assign_labels = 'kmeans'	No
<i>Agglomerative Clustering</i>	sklearn.cluster	AgglomerativeClustering	K, metric = 'euclidean/cosine', linkage = 'ward/average'*	No
<i>CLARA</i>	Scikit-learn-extra	CLARA	K, metric = 'correlation'/'braycurtis'**	Yes
<i>FCM</i>	fuzzy_c_means. fcmeans	FCM	K	Yes
<i>GMM</i>	sklearn.mixture	GaussianMixture	K, covariance_type = 'diag'	Yes
<i>BIRCH</i>	sklearn.cluster	Birch	K	No
<i>SOM</i>	sklearn_som	SOM	m=2, n=1, dim=46/18***	Yes
<i>Mean-Shift</i>	sklearn.cluster	Mean_Shift	bin_seeding = True, cluster_all = True, min_bin_freq = 1/2/3****	No

*: the 'Euclidean/Ward' and 'Cosine/Average' parameters referred to ACS and ischemia database, respectively.

**: the 'correlation/braycurtis' distance metrics referred to ACS and ischemia database, respectively.

***: 'dim' indicates the dimensionality of the input space, so it assumes 46 and 18 according to the number of features considered.

****: 'min_bin_freq' specifies the minimum number of points that a bin must contain to be considered for the mean shift procedure and assumes 1 considering all features, while 2 and 3 with ischemia and ACS database composed by 18 features. These values are chosen to obtain two final clusters.

Next, each clustering method applied to the training dataset to build and learn the model. This is achieved by using '*fit*' method to obtain the fitted instance and then by '*labels*' attribute to retrieve the cluster labels. These labels are then assigned meanings of 'pathological' and 'normal' based on the abundance criterion, utilizing the '*counter*' function for quantitatively assessing the sizes of the clusters. Specifically, the largest cluster is assigned the meaning of 'normal', while the

less populous cluster belongs to 'pathological', based on the prevalence of health conditions within a healthcare database. The trained model is subsequently applied to the testing set to provide an unbiased evaluation of the final model's performance, utilizing the '*predict*' method for most algorithms. However, Agglomerative and Spectral clustering methods do not have this method for classifying new data after training, as they identify the structure of clusters but lack an explicit mechanism for assigning new data points to these clusters. To achieve cluster assignment on the testing data while maintaining an unsupervised approach, it was followed a similarity-based assignment method. This involved calculating the distance and affinity matrices for testing data using the same kernel function applied to the training data, as specified in *Table VII*, for Agglomerative and Spectral clustering, respectively. This approach was useful for identifying the training set points that are most similar to each point in the testing set based on the characteristics of the matrices, allowing for the assignment of testing labels.

4.3 Statistical Analysis

Finally, a statistical analysis was performed to help providers make informed clinical decisions and guide patient care by understanding the likelihood of a patient having a disease through diagnostic tests that assess the appropriateness of the diagnostic tool. A gold standard is a set of correct annotations or correct answers to a query or the correct classification of documents. For the specific analysis corresponded to the target, so 0 and 1. It is used to evaluate the performance of any method, quantified in terms of its Sensitivity (SE), Precision (PR), Specificity (SP), Accuracy (ACC) and F1-score (F1), which are computed according *Eq.s (20) (21) (22) (23) (24)*, respectively,

$$SE = 100 \times \frac{TP}{TP+FN} \quad (20)$$

$$SP = 100 \times \frac{TN}{TN+FP} \quad (21)$$

$$PR = 100 \times \frac{TP}{TP+FP} \quad (22)$$

$$ACC = 100 \times \frac{TP+TN}{(TP+FP+TN+FN)} \quad (23)$$

$$F1 = 2 \times \frac{PR \times SE}{PR + SE} \quad (24)$$

where TP stands true positive, FP for false positive, TN for true negative and FN for false negative, respectively [72]. If an ischemic beat is classified as the ischemic, then it is said to be TP, while if it is a normal beat, it is said to be TN; any normal beat classified as an ischemic beat by mistake then it said to be a FP, similarly if any ischemic beat which is classified as normal beat by mistake it is said to be a FN. The SE of a test is its ability to determine the patient cases correctly, thus it estimates by calculating the proportion of TP in-patient cases. Whereas the SP of a test is its ability to determine the healthy cases correctly, so it obtained by the proportion of TN in healthy cases. PR identifies how accurately the model predicted the positive classes, where the number of true positive events divided by the sum of positive true and false events. The ACC of a test is its ability to differentiate the patient and healthy cases correctly. Its estimate depends on both the proportion of TP and TN in all evaluated cases [73], as it is defined as a weighted arithmetic mean of precision and inverse precision. ACC can also be high but precision low, meaning the system performs well but the results produced are slightly spread [74]. F1 score means the harmonic mean between PR and recall, that coincides with SE. F1 obtained by imposing the weight function $\beta=1$ in F-score definition, reported in [74]. F1-score is a composite measure which benefits algorithms with higher sensitivity and challenges algorithms with higher specificity [75], since it balances precision and recall in the positive class, while accuracy looks at correctly classified observations, both positive and negative. That makes a big difference, especially for the imbalanced problems, where by default the model will be good at predicting true negatives and hence accuracy will be high. It provides a more informative measure than ACC for problems where one class is much more frequent than the other, because it considers both false positives and false negatives. TP, FP, TN, and FN derived from the Confusion Matrix (CM), as shown in Eq. (25). It was generated using the function ‘*confusion_matrix*’, which takes as input the true labels, corresponding to the gold standard, and predicted labels from the model. It is a table, named *CM*, with columns containing actual classes and the rows with predicted classes, and it describes the classifier's performance against the known test data.

$$CM = \begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix} \quad (25)$$

Chapter 5 – Results

The performance metrics described in *Chapter 4*, were computed for both the databases, considering all features and then limiting to the first 18. This analysis aims to evaluate the effectiveness of the models in predicting outcomes based on the selected features, particularly the role of serial ones in disease detection in combination with direct measurements. The outcomes of the statistical analysis conducted in this study presented in *Tables VIII, IX, X, XI*. The latter serve to illustrate and compare the performance of various clustering methods that were applied to both the training and testing datasets, to which are allocated 70% and 30% of the data, respectively. This division is crucial, as it allows for a robust evaluation of the clustering methods' effectiveness in identifying patterns and structures within the data. Each table shows the comparative analysis of the clustering techniques, allowing for a comprehensive understanding of their strengths and weaknesses in relation to the dataset, employed in the specific case. The tables highlight the highest values of each index, compared to the database and the number of features considered, using bold font.

Table VIII – Evaluation metrics used for measuring the performance of the myocardial ischemia database, considering only the direct AECG measurements. SE: sensitivity. SP: specificity. ACC: accuracy. F1: F1-score.

METHODS	Training (70%)				Testing (30%)			
	<i>SE (%)</i>	<i>SP (%)</i>	<i>ACC (%)</i>	<i>F1 (%)</i>	<i>SE (%)</i>	<i>SP (%)</i>	<i>ACC (%)</i>	<i>F1 (%)</i>
<i>K-Means</i>	30,51	93,37	84,40	45,44	41,18	93,42	85,92	56,59
<i>K-Medoids</i>	66,95	82,79	80,53	78,12	68,63	81,25	79,44	79,30
<i>Spectral Clustering</i>	86,04	57,63	81,98	55,28	85,20	64,71	82,25	56,54
<i>Agglomerative Clustering</i>	95,91	34,75	87,18	72,73	97,70	45,10	90,14	85,91
<i>CLARA</i>	64,41	79,69	77,51	76,13	66,67	76,32	74,93	77,72
<i>FCM</i>	89,00	42,37	82,35	54,29	85,86	52,94	81,13	53,23
<i>GMM</i>	58,47	80,39	77,27	71,53	66,67	78,62	76,90	77,79
<i>BIRCH</i>	89,84	41,52	82,95	55,83	89,14	49,02	83,38	58,11
<i>SOM</i>	80,96	64,41	78,60	49,86	77,3	64,71	75,49	45,61
<i>Mean Shift</i>	16,95	98,31	86,70	28,41	27,45	99,67	89,3	41,97

Table IX – Evaluation metrics used for measuring the performance of the myocardial ischemia database, considering all the features, both direct AECG and serial AECG-RECG. SE: sensitivity. SP: specificity. ACC: accuracy. F1: F1-score.

METHODS	Training (70%)				Testing (30%)			
	<i>SE (%)</i>	<i>SP (%)</i>	<i>ACC (%)</i>	<i>F1 (%)</i>	<i>SE (%)</i>	<i>SP (%)</i>	<i>ACC (%)</i>	<i>F1 (%)</i>
<i>K-Means</i>	95,20	33,05	86,34	68,44	96,38	43,14	88,73	78,82
<i>K-Medoids</i>	81,38	69,49	79,69	52,10	81,25	76,47	80,56	54,17
<i>Spectral Clustering</i>	86,88	65,25	83,80	59,55	85,20	66,67	82,53	57,19
<i>Agglomerative Clustering</i>	37,29	97,74	89,12	52,79	50,98	98,03	91,27	65,67
<i>CLARA</i>	71,19	78,7	77,63	81,11	70,59	75,33	74,65	80,57
<i>FCM</i>	90,41	49,15	84,52	61,00	88,49	60,78	84,51	61,37
<i>GMM</i>	79,13	62,71	76,78	46,91	77,63	60,78	75,21	44,63
<i>BIRCH</i>	97,04	30,51	87,54	76,52	97,70	43,14	89,86	85,41
<i>SOM</i>	60,17	86,60	82,83	73,03	66,67	83,55	81,13	77,91
<i>Mean Shift</i>	31,36	96,61	87,30	46,43	39,22	97,70	89,30	54,73

Table X – Evaluation metrics used for measuring the performance of the ACS database, considering only the direct AECG measurements. SE: sensitivity. SP: specificity. ACC: accuracy. F1: F1-score.

METHODS	Training (70%)				Testing (30%)			
	<i>SE (%)</i>	<i>SP (%)</i>	<i>ACC (%)</i>	<i>F1 (%)</i>	<i>SE (%)</i>	<i>SP (%)</i>	<i>ACC (%)</i>	<i>F1 (%)</i>
<i>K-Means</i>	89,74	26,85	80,34	46,63	89,01	25,00	79,44	43,26
<i>K-Medoids</i>	44,29	79,48	74,22	59,16	56,25	81,87	78,04	69,64
<i>Spectral Clustering</i>	42,28	81,37	75,53	57,31	51,56	82,69	78,04	65,74
<i>Agglomerative Clustering</i>	86,08	32,21	78,03	43,29	84,89	35,94	77,57	43,77
<i>CLARA</i>	46,98	76,18	71,82	61,52	59,38	78,30	75,47	72,06
<i>FCM</i>	34,23	83,73	76,33	49,27	34,38	81,59	74,53	49,38
<i>GMM</i>	75,47	44,30	70,81	36,52	76,10	45,31	71,50	37,64
<i>BIRCH</i>	86,08	32,21	78,03	43,29	84,89	35,94	77,57	43,77
<i>SOM</i>	76,89	44,30	72,02	37,95	75,00	43,75	70,33	35,82
<i>Mean Shift</i>	11,41	98,00	85,06	20,15	20,31	97,25	85,75	32,96

Table XI – Evaluation metrics used for measuring the performance of the ACS database, considering all the features, both direct AECG and serial AECG-RECG. SE: sensitivity. SP: specificity. ACC: accuracy. F1: F1-score.

METHODS	Training (70%)				Testing (30%)			
	SE (%)	SP (%)	ACC (%)	F1 (%)	SE (%)	SP (%)	ACC (%)	F1 (%)
<i>K-Means</i>	92,10	25,50	82,15	51,96	92,58	26,56	82,71	54,52
<i>K-Medoids</i>	78,66	46,98	73,92	41,18	79,94	59,38	76,87	47,94
<i>Spectral Clustering</i>	80,42	44,30	75,02	42,03	83,24	54,69	78,97	50,71
<i>Agglomerative Clustering</i>	88,56	28,19	79,54	45,06	87,91	25,00	78,50	40,92
<i>CLARA</i>	44,30	83,14	77,33	59,25	56,25	85,71	81,31	69,75
<i>FCM</i>	38,26	84,67	77,73	53,44	42,19	83,79	77,57	57,29
<i>GMM</i>	73,82	44,30	69,41	34,98	74,45	51,56	71,03	38,75
<i>BIRCH</i>	88,56	28,19	79,54	45,06	87,91	25,00	78,50	40,92
<i>SOM</i>	81,84	42,28	75,93	42,86	80,49	45,31	75,23	42,64
<i>Mean Shift</i>	17,45	95,4	83,75	29,06	23,44	95,33	84,58	36,98

Fig.s 11 and 12, created in Microsoft Excel, are the Line Charts that display the trends of the F1 score for both databases, illustrating how the F1 score varies with different feature sets. This allows for evaluating which features are necessary and how the performance of clustering methods changes with different feature spaces as input, due to the imbalanced data.

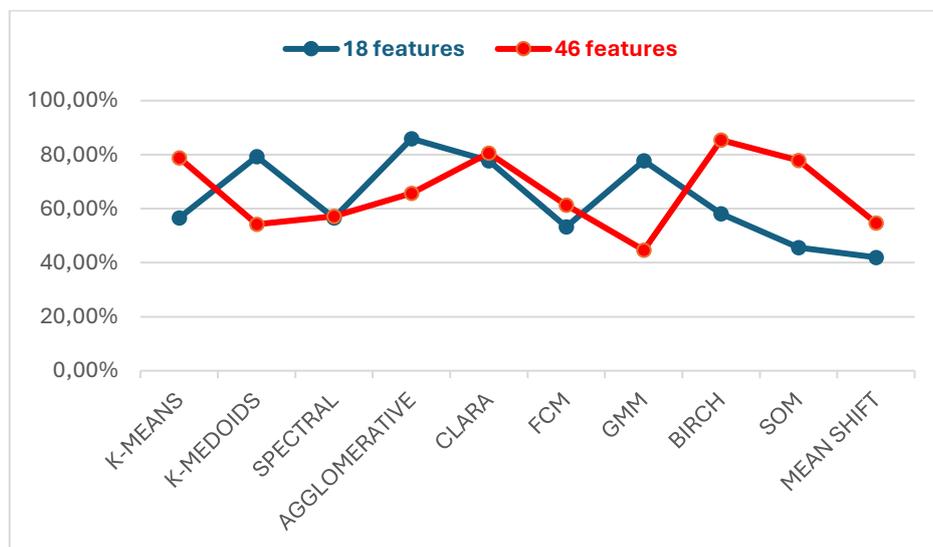


Fig. 11 – Line Chart of F1 score in Myocardial Ischemia database.

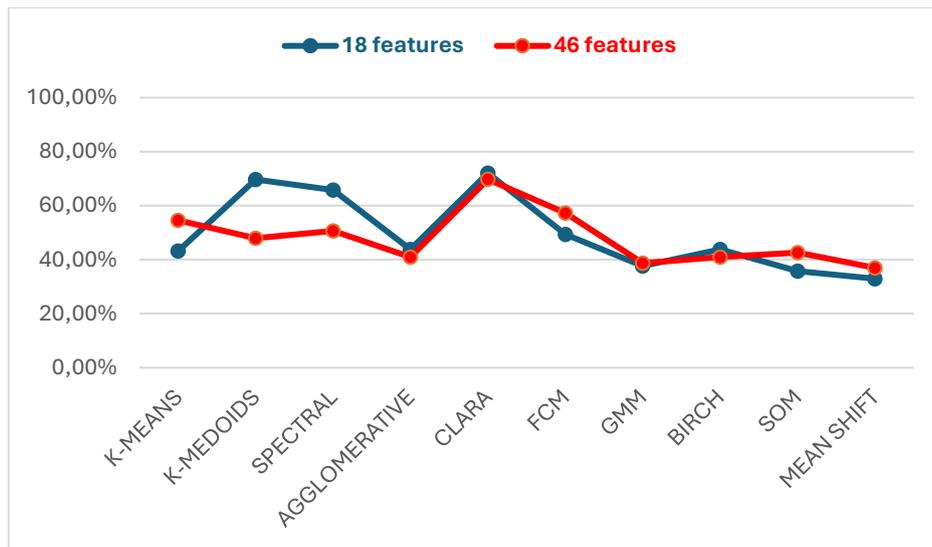


Fig. 12 – Line Chart of F1 score in Acute Coronary Syndrome database.

Overall, Agglomerative Clustering results in the best clustering method in the analysis of myocardial ischemia database, summarized in *Table VIII and IX*, considering the highest ACC in both the studied cases, equal to 90,14% and 91,27% passing from 18 to 46 features. Specifically, it demonstrates superior SP when utilizing all the features and the best SE with the first 18. The positive assessment of Agglomerative Clustering is further supported by the F1 score, which assumes 85,91%, the highest value, when considering only the direct measurements. Whereas, with all the features, the best methods are CLARA and BIRCH, obtaining 80,57 and 85,41%, respectively. Turning to the ACS database, for which results are reported on *Table X and XI*, in both the feature sets the best SE is obtained with the K-means, while Mean Shift excels in providing the best SP and ACC. A common outcome between the two databases is the overall higher F1 score obtained in CLARA. In analyzing which feature set is more effective between the two considered, it is possible to conclude that the ACS database reveals a minimal difference in performance, especially in terms of F1 as shown in *Fig. 12*. Conversely, when examining the myocardial ischemia database, this conclusion holds true for most clustering methods, but not for all, as observed in *Fig. 11*.

Chapter 6 – Discussion

As the demand for rapid and accurate cardiac assessment increases in emergency medical settings, the integration of advanced ML methods may present a promising approach to enhance and speed up diagnostic accuracy. The challenges emerge from to the large-scale and high-speed interactions, occurring in various contexts, including healthcare, where inaccurate diagnoses can significantly impact patient risk. To address these issues, unsupervised ML can be employed for discovering data distributions from large datasets, learning the underlying structures of data, facilitating decision-making in complex environments, without any predefined labels. In the analysis of clustering methods applied to the myocardial ischemia and ACS databases, the aim is to demonstrate their effectiveness and performance as a supportive tool in pre-hospital electrocardiography for enhancing diagnostic accuracy, streamline workflow, and ultimately improving patient outcomes. The choice of the parameters is particularly important, as it can significantly influence the results, especially with imbalanced datasets, where the minority class, consisting of pathological subjects, may be not well identified from the majority class, leading to potential misclassifications. Therefore, careful tuning of parameters, such as the number of clusters, type of linkage and distance metrics, is essential to improve the model's ability to accurately identify and predict the minority class.

Looking at *Tables VIII and IX*, Agglomerative Clustering emerges as the most effective technique, comprehensively. This conclusion derives from the high values assumed by ACC and F1 when considering both feature sets. Particularly, F1 of 85,91% is the highest one when analyzing only the direct AECG measurements. In contrast, when all features are taken into account, alternative clustering methods, as CLARA and BIRCH, demonstrate superior performance, suggesting that these methods may be better suited for more complex and large feature sets. Notably, Agglomerative Clustering exhibits also superior SP, equal to 98,03%, when utilizing all available features, which underscores its capability to effectively identify healthy subjects within the complete database. Furthermore, it achieves quite the same score for SE when considering the first 18 features, indicating its effectiveness in correctly identifying controls with a reduced amount of data. The reliable performance may derive from the choice of the *average* and *ward* as linkages, which are generally better suited for scenarios where the data is not uniformly distributed. Indeed, they help in minimizing the impact of outliers and creating clusters that are well-separated and more representative of the underlying data structure, differently from the other types, listed in *Table IV*. Whereas Mean Shift method with SP of 99,67% when considering the reduced feature

set is able to detect basically all TN, avoiding them unnecessary treatments. Overall, an high performance in identifying TP is also achieved by Spectral Clustering and FCM, while K-Medoids performs well in terms of TN identification. The ACC remains high for all methods, as it does not fall below 74,65%, meaning that the models are performing reasonably well, as they are correctly classifying more than two-thirds of the patients.

Passing to the ACS database, the results, reported on *Tables X and XI*, reveal a quite similar hierarchy of clustering effectiveness, although with a slight overall decrease in performance. Before starting to discuss in-depth the findings, the reason found in the definition of the Gold Standard for the two syndromes. Indeed, in the case of myocardial ischemia it is well-defined and assessed in literature, differently from the ACS. Fractional flow reserve (FFR) is the current Gold Standard for invasive in the cath-lab assessment of ischemia in patients with stable CAD. FFR defined as the ratio of distal and proximal mean pressures across stenosis during maximal hyperemia. This confirmed by the fact that when maximal vasodilatation achieved, the correlation between coronary pressure and coronary flow becomes linear, allowing measurement of the flow reduction related to a coronary stenosis through the pressure drop across it [76]. For what regards the statistical analysis, K-means results the best clustering method in terms of SE, remaining above 89% in both datasets, demonstrating its reliability in accurately identifying positive cases, as for the other database when considering all features. Agglomerative Clustering and BIRCH also provide a good SE, over 84% in both feature sets, as also seen in myocardial ischemia database. Meanwhile, the Mean Shift algorithm achieves the highest SP and ACC, indicating its capability in detecting controls, as also highlighted previously. Interestingly, CLARA method yields the highest F1 in both analyses, where does not fall below 69,75%, proving to balance false positives and false negatives well and to be a strong classifier. This confirms that CLARA is a versatile clustering method, capable of delivering robust outcomes across different datasets and conditions. However, it is evident that SP assumes lower values than the first database studied, except for FCM and CLARA. This can be justified by the assumptions and characteristics of each method. A possible key difference stays on the flexibility of assigning the corresponding cluster to the data point, which differs between soft and hard methodologies, as discussed in *Section 3.3*. Hard clustering methods, such as K-means, Agglomerative clustering, and BIRCH, assign each data point to a single cluster. Whereas soft clustering methods, like FCM, allow data points to belong to multiple clusters with varying degrees of membership. This flexibility can influence the overall performance metrics, including SE and SP. Although, ACC values are slightly lower than the first database, with a minimum score of 70,33%, which is still considered a valuable outcome.

Interestingly, the evaluation of the role of direct measurements in disease detection can be assessed by analyzing the differences in clustering performance, measured in terms of ACC and F1, by varying the feature set. This disparity exhibits quite a common trend across the two studied databases. The analysis conducted on ACS database shows a minimal variation in performance, particularly concerning F1, as illustrated in *Fig. 12*. This suggests that for this condition, it is sufficient to provide only the 18 direct measurements for making the diagnosis. On the other hand, when examining the myocardial ischemia database, this finding remains true for the majority of clustering methods, but not for all. Indeed, K-means, BIRCH, SOM, and Mean Shift show an improvement in performance of F1, when the feature set enlarged from 18 to 46 features. Thus, these algorithms benefit from the inclusion of serial features, enhancing their ability to identify meaningful and well-separated clusters in high-dimensional spaces. Conversely, K-medoids and GMM algorithms demonstrated a decline in clustering performance as the dimensionality increased, meaning they are more effective with the first 18 features. The other clustering techniques do not highlight any significant performance variation when increasing the number of features, demonstrating their uncorrelation between feature dimensionality and clustering performance, too. These outcomes highlight the importance of considering the dimensionality of the data when choosing the appropriate clustering methodology, as different algorithms respond uniquely to changes in feature space. Furthermore, this finding reveals that the serial features are not fundamental for reaching a final diagnosis, meaning that the availability of the RECG is not essential for an accurate and prompt diagnosis, even because it is often difficult to retrieve in most of the cases.

As anticipated, a comparison of the evaluation metrics of testing sets between databases of same size confirms that the algorithms perform significantly better when applied to the myocardial ischemia dataset. This conclusion supported by the observed higher values of ACC and F1 metrics, especially the last one, which is notably twice as high in the myocardial ischemia database compared to ACS one. This trend is evident in the results of Agglomerative Clustering and GMM when utilizing 18 features, as well as in BIRCH and SOM with 46 features. Additionally, the difference is also marked when using all data with K-means and Agglomerative Clustering algorithms. This indicates that these algorithms are not only more effective with the myocardial ischemia data, but it also suggests that the underlying characteristics of the dataset may lead to better clustering and classification outcomes, influencing the overall algorithm performance.

Finally, all algorithms implemented in Google Colab demonstrate comparable execution speeds and time consumption, resulting valuable tools in ED settings, where timely decision-making is crucial. By ensuring rapid processing, they can support healthcare professionals in quickly analyzing medical reports and making diagnoses, thereby reducing the risk of health complications, a primary concern for every physician.

In summary, the comparative analysis of clustering methods reveals nuanced strengths and weaknesses, with Agglomerative clustering and CLARA emerging as particularly notable clustering methods in the context of the myocardial ischemia database, while CLARA and K-medoids demonstrate their respective advantages in the ACS. Regarding the evaluation metrics, ACC and F1 assume the highest value in Mean Shift and CLARA, with a minimum score of 84,58% and 69,75%, respectively. However, CLARA exhibits the overall highest performance, resulting in the best method in terms of robustness against outliers, effectiveness with imbalanced data, adaptability to different database characteristics, and cases detection.

The comparative analysis conducted in this study also highlights the limitations of considered unsupervised ML techniques relative to traditional diagnostic methods. While unsupervised ML emerges as a promising tool in classifying pathological subjects from healthy individuals, a significant challenge remains establishing clear criteria for interpreting the outputs of clustering models. Indeed, in this study, the abundance criterion used to assign labels to the identified cluster, consisting of designing the most populous cluster as representing cases and labeling the remaining as controls. While this approach gives positive outcomes, it is essential to explore and develop alternative criteria, which may be more widely applicable and capable of enhancing the performance of these innovative methods in future research. An additional limitation of this work is the choice to implement these techniques only in Python. While it is a powerful language, there may be other programming languages that offer more extensive functionality and potentially provide better performance. Finally, this study uses the tuning parameters to select the combinations of those giving the highest ACC. Although this approach demonstrates to be effective for the thesis purposes, future research may explore alternative strategies for parameter selection and output optimization.

Chapter 7 – Conclusion

This thesis explores the potential of unsupervised ML as a supportive tool in the context of ED. There are implemented and analyzed a total of ten unsupervised ML techniques with the aim to evaluate their performance in detecting cases in myocardial ischemia and ACS databases, utilizing only direct AECG measurements and the union of the previous with serial AECG-RECG features for each subject, who participated to the SUBTRACT study. Moreover, this work attempts to determine the differing roles of these two classes of feature set in the diagnosis of the pathologies under study in a pre-hospital setting. Through a comprehensive evaluation of performance metrics, it emerged that: (1) CLARA resulted compressively the best clustering method across both databases and feature sets; since ACC and F1 do not fall below 74,65% and 69,75%, respectively; (2) in ACS database the serial features were found to be irrelevant for the diseases detection, differently from the myocardial ischemia, where some algorithms benefited from the inclusion of serial features.

One potential weakness characterizing the almost of unsupervised ML methods regards the requirement of specifying in input the number of clusters, which may be not so obvious to know in advance in practical applications, especially in fields like healthcare. An incorrect choice can lead to misleading results or an inaccurate representation of the underlying data. For instance, there may exist for the same pathological condition multiple degrees of severity, requiring different therapies and, if not recognized in a timely manner, causing diverse severe complications. Furthermore, the clinical application of some of the unsupervised techniques in the real clinical scenario depends upon the availability of a prior ECG of the same patient, used as a reference. This limitation may be solved by introducing a unique, accessible cloud storage service dedicated to ECG recordings within each region. By addressing these procedural issues, it can be thoughtful the integration of advanced ML methods into routine diagnostic processes of ECG recordings in pre-hospital care processes, ultimately improving patient outcomes.

References

- [1] Sbröllini A, Haar C C t, Leoni C, Morettini M, Burattini L, Swenne C A, “*Advanced repeated structuring and learning procedure to detect acute myocardial ischemia in serial 12-lead ECGs*”, *Physiological Measurement*, Volume 44, Number 8, 2023, <https://doi.org/10.1088/1361-6579/ace241>.
- [2] Roth G A, Mensah G A, Johnson C O, Addolorato G, Ammirati E, Baddour L M, Barengo N C, Beaton A Z B, Benjamin E J, Benziger C, Bonny A, Brauer M, Brodmann M, Cahill T J, Carapetis J, Catapano A L, Chugh S S, Cooper L T, Coresh J, Criqui M, Fuster V, “*Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study*”, *Journal of the American College of Cardiology*, 76(25):2982-3021, 2020, <https://doi.org/10.1016/j.jacc.2020.11.010>.
- [3] Santos-Gallego C G, Picatoste B, Badimón J J, “*Pathophysiology of Acute Coronary Syndrome*”, *Springer Science+Business Media*, 16:401, 2014, <https://doi.org/10.1007/s11883-014-0401-9>.
- [4] Byrne R A, Coughlan J J, Barbato E, Berry C, Chieffo A, Claeys M J, Dan G A, Dweck M R, Galbraith M, Martine Gilard, Lynne Hinterbuchner, Jankowska E A, Jüni P, Kimura T, Kunadian V, Leosdottir M, Lorusso R, Pedretti R F E, Rigopoulos A G, Gimenez M R, Thiele H, Vranckx P, Wassmann S, Wenger N K, Ibanez B, “*2023 ESC Guidelines for the management of acute coronary syndromes: Developed by the task force on the management of acute coronary syndromes of the European Society of Cardiology (ESC)*”, *European Heart Journal*, 44(38):3720–3826, 2023, <https://doi.org/10.1093/eurheartj/ehad191>.
- [5] Khan M A, Hashim M J, Mustafa H, Baniyas M Y, Al Suwaidi S K, AlKatheeri R, Alblooshi F M K, Almatrooshi M E, Alzaabi M E, Al Darmaki R S, Lootah S N, “*Global Epidemiology of Ischemic Heart Disease: Results from the Global Burden of Disease Study*”, *Cureus*, 12(7):e9349, 2020, <https://doi.org/10.7759/cureus.9349>.
- [6] Benjamin E J, Blaha M J, Chiuve S E, Cushman M, Das S R, Deo R, Muntner P, “*Heart disease and stroke statistics-2017 update: A report from the American Heart Association*”, *Circulation*, 135: e146–e603, 2017, <https://doi.org/10.1161/cir.0000000000000485>.
- [7] Clifford G D, Azuaje F, McSharry P E, *Advanced Methods and Tools for ECG Data Analysis*, ARTECH HOUSE, 2006.
- [8] Severino P, D’Amato A, Pucci M, Infusino F, Adamo F, Birtolo L I, Netti L, Montefusco G, Chimenti C, Lavallo C, Maestrini V, Mancone M, Chilian W M, Fedele F, “*Ischemic Heart Disease Pathophysiology Paradigms Overview: From Plaque Activation to Microvascular Dysfunction*”, *International Journal of Molecular Sciences*, 21, 2020, <https://doi.org/10.3390/ijms21218118>.

- [9] Farthing D E, Farthing C A, Xi L, “*Inosine and hypoxanthine as novel biomarkers for cardiac ischemia: From bench to point-of-care*”, *Experimental Biology and Medicine*, 240: 821–831, 2015, <https://doi.org/10.1177/1535370215584931>.
- [10] Pagliaro B R, Cannata F, Stefanini G G, Bolognese L, “*Myocardial ischemia and coronary disease in heart failure*”, *Springer Science+Business Media*, 25: 53-65, 2019, <https://doi.org/10.1007/s10741-019-09831-z>.
- [11] Kaski J C, Crea F, Gersh B J, Camici P G, “*Reappraisal of Ischemic Heart Disease Fundamental Role of Coronary Microvascular Dysfunction in the Pathogenesis of Angina Pectoris*”, *Circulation*, 138: 1463-1480, 2018, <https://doi.org/10.1161/CIRCULATIONAHA.118.031373>.
- [12] Ganapathy N, Swaminathan R, Deserno T M, “*Deep Learning on 1-D Biosignals: a Taxonomy-based Survey*”, *IMIA Yearbook of Medical Informatics*, 27(1): 98-109, 2018, <http://dx.doi.org/10.1055/s-0038-1667083>.
- [13] Manocha A M, Singh M, “*An Overview of Ischemia Detection Techniques, International Journal of Scientific & Engineering Research*”, Volume 2, Issue 11, 2011.
- [14] Kihlgren M, Almqvist C, Amankhani F, Jonasson L, Norman C, Perez M, Ebrahimi A, Gottfridsson C, “*The U-wave: A remaining enigma of the electrocardiogram, Journal of Electrocardiology*”, 79: 13-20, 2023, <https://doi.org/10.1016/j.jelectrocard.2023.03.001>.
- [15] Winkler C, Funk M, Schindler D M, “*Arrhythmias in patients with acute coronary syndrome in the first 24 hours of hospitalization*”, *Heart & Lung*, 42(6): 422–7, 2013, <https://doi.org/10.1016/j.hrtlng.2013.07.010>.
- [16] Zègre-Hemsey J K, Garvey J L, Carey M G, “*Cardiac Monitoring in the Emergency Department*”, *Critical Care Nursing Clinics of North America*, 28(3): 331–345, 2016, <https://doi.org/10.1016/j.cnc.2016.04.009>.
- [17] Marzilli M, Crea F, Morrone D, Bonow R O, Brown D L, Camici P G, Chilian W M, DeMaria A, Guarini G, Noel Bairey Merz A H J, Pepine C, Scali M S, Weintraub W S, Boden W E, “*Myocardial ischemia - From disease to syndrome*”, *International Journal of Cardiology*, 314: 32–35, 2020, <https://doi.org/10.1016/j.ijcard.2020.04.074>.
- [18] Smit M, Coetsee A R, Lochner A, “*The Pathophysiology of Myocardial Ischemia and Perioperative Myocardial Infarction*”, *Journal of Cardiothoracic and Vascular Anesthesia*, 34:2501-2512, 2020, <https://doi.org/10.1053/j.jvca.2019.10.005>.
- [19] Cohen D J, Foley R W, Ryan J M, “*Intraoperative coronary artery spasm successfully treated with nitroglycerin and nifedipine*”, *The Annals of Thoracic Surgery*, 36: 97-100, 1983.

- [20] Buxton A E, Goldberg S, Harken A, “*Coronary-artery spasm immediately after myocardial revascularization*”, *The New England Journal of Medicine*, 304:1249-53, 1981.
- [21] Swenne C A, Haar C C t, “*Context-independent identification of myocardial ischemia in the prehospital ECG of chest pain patients*”, *Journal of Electrocardiology*, 82:34-41, 2024, <https://doi.org/10.1016/j.jelectrocard.2023.10.009>.
- [22] Monizzi G, Di Lenarda F, Gallinoro E, Bartorelli A L, “*Myocardial Ischemia: Differentiating between Epicardial Coronary Artery Atherosclerosis, Microvascular Dysfunction and Vasospasm in the Catheterization Laboratory*”, *Journal of Clinical Medicine*, 13: 4172, 2024, <https://doi.org/10.3390/jcm13144172>.
- [23] Shimokawa H, Yasuda S, “*Myocardial ischemia: Current concepts and future perspectives*”, *Journal of Cardiology*, 52: 67-78, 2008, <https://doi.org/10.1016/j.jjcc.2008.07.016>.
- [24] Barstow C, Rice M, McDivitt J D, “*Acute Coronary Syndrome, Diagnostic Evaluation*”, *American Family Physician*, 95(3): 170-177, 2017.
- [25] Ibanez B, “*2017 ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation: The Task Force for the management of acute myocardial infarction in patients presenting with ST-segment elevation of the European Society of Cardiology (ESC)*”, *European Heart Journal*, 39: 119-77, 2017.
- [26] Smith J N, Negrelli J M, Manek M B, Hawes E M, Viera A J, “*Diagnosis and Management of Acute Coronary Syndrome: An Evidence-Based Update*”, *American Board of Family Medicine*, Volume 28, Number 2, 2015, <https://doi.org/10.3122/jabfm.2015.02.140189>.
- [27] Birnbaum Y, Wilson J M, Fiol M, de Luna A B, Eskola M, Nikus K, “*ECG Diagnosis and Classification of Acute Coronary Syndromes*”, *Annals of Noninvasive Electrocardiology*, 19(1): 4-14, 2014, <https://doi.org/10.1111/anec.12130>.
- [28] Crea F, Libby P, “*Acute Coronary Syndromes: The Way Forward From Mechanisms to Precision Treatment*”, *Circulation*, 136: 1155-1166, 2017, <https://doi.org/10.1161/CIRCULATIONAHA.117.029870>.
- [29] Kumar A, Cannon C P, “*Acute Coronary Syndromes: Diagnosis and Management, Part I*”; *Mayo Clinic Proceedings*, 84(10): 917-938, 2009, <https://doi.org/10.4065/84.10.917>.
- [30] Jneid H, Anderson J L, Wright R S, Adams C D, Bridges C R, Casey D R, Ettinger S M, Fesmire F M, Ganiats T G, Lincoff A M, Peterson E D, Philippides G J, Theroux P, Wenger N K, Zidar J P, “*2012 ACCF/AHA Focused Update Incorporated Into the ACCF/AHA 2007 Guidelines for the Management of Patients With Unstable Angina/Non-ST-Elevation Myocardial Infarction. A Report of the American*

College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines”, *Circulation*, 127: e663-e828, 2012, <https://doi.org/10.1161/CIR.0b013e31828478ac>.

[31] Falk E, Nakano M, Bentzon J F, Finn A V, Virmani R, “*Update on acute coronary syndromes: the pathologists’ view*”, *European Heart Journal*, 34: 719-28, 2013.

[32] Achar S J A, Kundu S, Norcross W A, “*Diagnosis of Acute Coronary Syndrome, American Family Physician*”, 72(1): 119-26, 2005.

[33] Amsterdam E A, Kirk J D, Bluemke D A, “*Testing of low-risk patients presenting to the emergency department with chest pain: a scientific statement from the American Heart Association*”, *Circulation*, 122(17): 1756-76, 2010, <https://doi.org/10.1161/CIR.0b013e3181ec61df>.

[34] Hess C N, Wang T Y, McCoy L A, Messenger J C, Effron M B, Zettler M E, Fonarow G C, “*Unplanned inpatient and observation rehospitalizations after acute myocardial infarction: Insights from the TRANSLATE-ACS study*”, *Circulation*, 133: 493–501, 2015, <https://doi.org/10.1161/circulationaha.115.017001>.

[35] Lee T H, Cook E F, Weisberg M, Sargent R K, Wilson C, Goldman L, “*Acute chest pain in the emergency room: identification and examination of low-risk patients*”, *Archives of Internal Medicine*, 145: 65–9, 1985.

[36] Zègre-Hemsey J K, Burke L A, DeVon H A, “*Patient-reported symptoms improve prediction of acute coronary syndrome in the emergency department*”, *Research in Nursing and Health*, 41(5): 459-468, 2018, <https://doi.org/10.1002/nur.21902>.

[37] Haar C C t, Peters R J G, Bosch J, Sbröllini A, Gripenstedt S, Adams R, Bleijenberg E, Kirchhof C J H J, Dehnavi R A, Burattini L, de Winter R J, Macfarlane P W, Postema P G, Man S, Scherptong R W C, Schalij M J, Maan A C, Swenne C A, “*An initial exploration of subtraction electrocardiography to detect myocardial ischemia in the prehospital setting*”, *Annals of Noninvasive Electrocardiology*, 25:e12722, 2020, <https://doi.org/10.1111/anec.12722>.

[38] Kaul S, Ito H, “*Microvasculature in Acute Myocardial Ischemia: Part II. Evolving Concepts in Pathophysiology, Diagnosis, and Treatment, Clinical Cardiology: New Frontiers*”, *Circulation*, 109: 310-315, 2004, <https://doi.org/10.1161/01.CIR.0000111583.89777.F9>.

[39] Shen X, Lin C, Han L, Lin L, Pan L, Xiao-dong, “*Assessment of ischemia-modified albumin levels for emergency room diagnosis of acute coronary syndrome*”, *International Journal of Cardiology*, 149: 296-298, 2011, <https://doi.org/10.1016/j.ijcard.2010.01.013>.

[40] Nikus K, Pahlm O, Wagner G, Birnbaum Y, Cinca J, Clemmensen P, Eskola M, Fiol M, Goldwasser D, Gorgels A, Sclarovsky S, Stern S, Wellens H, Zareba W, de Luna A B, “*Electrocardiographic classification of acute coronary syndromes: a review by a committee of the International Society for Holter*

and Non-Invasive Electrocardiology”, Journal of Electrocardiology, 43: 91-103, 2010, <https://doi.org/10.1016/j.jelectrocard.2009.07.009>.

[41] Carmeliet E, “*Cardiac Ionic Currents and Acute Ischemia: From Channels to Arrhythmias*, *American Physiological Society*”, 79(3): 917-1017, <https://doi.org/10.1152/physrev.1999.79.3.917>.

[42] Wagner G S, “*Marriott’s Practical Electrocardiography*”, 9th ed., Baltimore, MD: Williams & Wilkins, 1994.

[43] Birnbaum Y, Nikus K, Kligfield P, Fiol M, Barrabés J A, Sionis A, Pahlm O, Niebla J G, de Luna A B, “*The Role of the ECG in Diagnosis, Risk Estimation, and Catheterization Laboratory Activation in Patients with Acute Coronary Syndromes: A Consensus Document*”, *Annals of Noninvasive Electrocardiology*, 19(5): 412-425, 2014, <https://doi.org/10.1111/anec.12196>.

[44] Drew B J, “*Practice Standards for Electrocardiographic Monitoring in Hospital Settings: An American Heart Association Scientific Statement from the Councils on Cardiovascular Nursing, Clinical Cardiology, and Cardiovascular Disease in the Young: Endorsed by the International Society of Computerized Electrocardiology and the American Association of Critical-Care Nurses*”, *Circulation*, 110(17): 2721-2746, 2004, <https://doi.org/10.1161/01.CIR.0000145144.56673.59>.

[45] Shusterman V, Goldberg A, Schindler D M, Fleischmann K E, Lux R L, Drew B J, “*Dynamic Tracking of Ischemia in the Surface Electrocardiogram*”, *Journal of Electrocardiology*, 40(6 Suppl): 179–186, 2008, <https://doi.org/10.1016/j.jelectrocard.2007.06.015>.

[46] Lloyd-Jones D M, Camargo C A, Lapuerta P, Giugliano R P, O’Donnell C J, “*Electrocardiographic and clinical predictors of acute myocardial infarction in patients with unstable angina pectoris*”, *American Journal of Cardiology*, 81: 1182-6, 1998, [https://doi.org/10.1016/s0002-9149\(98\)00155-6](https://doi.org/10.1016/s0002-9149(98)00155-6).

[47] Haar C C T, “*Difference vectors to describe dynamics of the ST segment and the ventricular gradient in acute ischemia*”, *Journal of Electrocardiology*, 46(4): 302–11, 2013, <https://doi.org/10.1016/j.jelectrocard.2013.04.004>.

[48] Komuro J, Kusumoto D, Hashimoto H, Yuasa S, “*Machine learning in cardiology: Clinical application and basic research*”, *Journal of Cardiology*, 82: 128-133, 2023, <https://doi.org/10.1016/j.jjcc.2023.04.020>.

[49] Ramesh A N, Kambhampati C, Monson J R T, Drew P J, “*Artificial intelligence in medicine*”, *The Annals of The Royal College of Surgeons of England*, 86: 334–8, 2004, <https://doi.org/10.1308/147870804290>.

[50] Maleki F, Ovens K, Najafian K, Forghani B, Reinhold C, Forghani R, “*Overview of Machine Learning Part 1. Fundamentals and Classic Approaches*”, *Neuroimaging Clinics of North America*, e17–e32, 2020, <https://doi.org/10.1016/j.nic.2020.08.007>.

- [51] Károly A I, Fullér R, Galambos P, “*Unsupervised Clustering for Deep Learning: A tutorial survey*”, Acta Polytechnica Hungarica, Volume 15, Number 8, 2018, <https://doi.org/10.12700/aph.15.8.2018.8.2>.
- [52] Deo R C, “*Machine Learning in Medicine*”, Circulation, 132: 1920-1930, 2015, <https://doi.org/10.1161/CIRCULATIONAHA.115.001593>.
- [53] Wei Wu, “*Unsupervised Learning*”, May 2022.
- [54] Singh J, “*A comprehensive review of clustering techniques in artificial intelligence for knowledge discovery: Taxonomy, challenges, applications and future prospects*”, Advanced Engineering Informatics, Volume 62, 2024 <https://doi.org/10.1016/j.aei.2024.102799>.
- [55] Grira N, Crucianu M, Boujemaa N, “*Unsupervised and Semi-supervised Clustering: a Brief Survey*”, Computer Science & Mathematics, 2005.
- [56] *Chapter 5: Measures of distance between samples: non-Euclidean.*
- [57] An Q, Rahman S, Zhou J, Kang J J, “*A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges*”, Sensors, Volume 23, 2023, <https://doi.org/10.3390/s23094178>.
- [58] Swarndeeep S J, Sharnil P, “*An Overview of Partitioning Algorithms in Clustering Techniques, International Journal of Advanced Research in Computer Engineering & Technology*”, Volume 5, Issue 6, 2016.
- [59] Chung F, “*Spectral graph theory*”, Conference Board of the Mathematical Sciences, Washington, Volume 92 (of the CBMS Regional Conference Series in Mathematics), 1997, <https://doi.org/10.1090/cbms/092>.
- [60] Mohar B, “*The Laplacian spectrum of graphs. In Graph theory, combinatorics, and applications*”, 2: 871-898, 1991.
- [61] Mohar B, “*Some applications of Laplace eigenvalues of graphs. In G. Hahn and G. Sabidussi (Eds.), Graph Symmetry: Algebraic Methods and Applications*”, 497: 225-275, 1997.
- [62] Luxburg U V, “*A Tutorial on Spectral Clustering*”, Springer, 17(4), 2007.
- [63] Nagpal A, Jatrain A, Gaur D, “*Review based on Data Clustering Algorithms*”, Conference on Information and Communication Technologies, 2013, <https://doi.org/10.1109/CICT.2013.6558109>.
- [64] Lucasius C B, Dane A D, Kateman G, “*On k-medoid clustering of large data sets with the aid of a genetic algorithm: background, feasibility and comparison*”, 282(3): 647-669, 1993 [https://doi.org/10.1016/0003-2670\(93\)80130-D](https://doi.org/10.1016/0003-2670(93)80130-D).

- [65] Abukmeil M, Ferrari S, Genovese A, Piuri V, Scotti F, “*A Survey of Unsupervised Generative Models for Exploratory Data Analysis and Representation Learning*”, ACM COMPUTING SURVEYS, Volume 1, Number 1, 2021, <https://doi.org/10.1145/3450963>.
- [66] Zhang T, Ramakrishnan R, Livny M, “*Birch: an efficient data clustering method for very large database*”, ACM SIGMOD Record, 25(2): 103-114, 1996, <https://doi.org/10.1145/235968.233324>.
- [67] Kohonen T, “*Self-Organized Formation of Topologically Correct Feature Maps*”, Springer – Biological Cybernetics, 43: 59-69, 1982.
- [68] Asan U, Ercan S, “*An Introduction to Self-Organizing Maps*”, Chapter 14, Atlantis Press Book, 2012, https://doi.org/10.2991/978-94-91216-77-0_14.
- [69] Cheng S S, Fu H C, Wang H M, “*Model-based clustering by probabilistic self-organizing maps*”, IEEE Transactions on Neural Networks and Learning, 20(5): 805-826, 2009. <https://doi.org/10.1109/TNN.2009.2013708>.
- [70] Demirovic D, “*An Implementation of the Mean Shift Algorithm, Image Processing On Line*”, 9: 251-268, 2019, <https://doi.org/10.5201/ipol.2019.255>.
- [71] Mason R E, Likar I, “*A new system of multiple-lead exercise electrocardiography*”, American Heart Journal, 71: 196-205, 1966, [https://doi.org/10.1016/0002-8703\(66\)90182-7](https://doi.org/10.1016/0002-8703(66)90182-7).
- [72] Shreffler J, Huecker M R, “*Diagnostic Testing Accuracy: Sensitivity, Specificity, Predictive Values and Likelihood Ratios*”, StatPearls Publishing; 2025, <https://www.ncbi.nlm.nih.gov/books/NBK557491/>.
- [73] Baratloo A, Hosseini M, Negida A, El Ashal G, “*Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity*”, Spring, 3(2): 48-49, 2015.
- [74] Dalianis H, “*Chapter 6: Evaluation Metrics and Evaluation*”, Springer, 2018, https://doi.org/10.1007/978-3-319-78503-5_6.
- [75] Sokolova M, Japkowicz N, Szpakowicz S, “*Beyond Accuracy, F-score and ROC: a Family of Discriminant Measures for Performance Evaluation*”, Springer, vol 4304, 2006, https://doi.org/10.1007/11941439_114.
- [76] Benenati S, De Maria G L, Scarsini R, Porto I, Banning A P, “*Invasive “in the cath-lab” assessment of myocardial ischemia in patients with coronary artery disease: When does the gold standard not apply?*”, Cardiovascular Revascularization Medicine, 19(3): 362-372, 2018, <https://doi.org/10.1016/j.carrev.2018.01.005>.