POLITECNICO DI TORINO

Master Degree in Biomedical Engineering

Master Thesis

Cardiovascular Risk Prediction in Rheumatic Patients by Artificial Intelligent Paradigms

---

Supervisor:

Prof. Marco Agostino DERIU

Co-supervisor:

Prof. Alberto AUDENINO

Author:

Michela SPERTI

Academic Year 2018-2019

"All models are wrong, but some of them are useful"

George E. P. Box

"Round numbers are always false"

Samuel Johnson

"Computers are incredibly fast, accurate and stupid.

Human Beings are incredibly slow, inaccurate and brilliant.

Together they are powerful beyond imagination."

Albert Einstein

Table of Contents

Abstract

According to World Health Organization, cardiovascular diseases are the first cause of death globally. People that present cardiovascular diseases or are at high cardiovascular risk (because of the presence of one or more risk factors such as hypertension, diabetes, hyperlipidemia or already established diseases) need early detection and preventive treatment. In this context, patients affected by inflammatory arthritis present an increased cardiovascular risk. In this field, cardiovascular diseases diagnosis is very tricky even for experts, due to the presence of many concurrent risk factors, some of which uncertain or unknown. Performances of traditional cardiovascular risk algorithms (such as Framingham, CUORE and SCORE) have already been assessed on patients with inflammatory arthritis, but the results show that they tend to underestimate the risk. For this reason, recently, the European League Against Rheumatism recommended to adapt the traditional algorithms with a multiplicative factor of 1.5 in patients with inflammatory arthritis. This work aims at exploring the use of the machine learning techniques to predict cardiovascular risk on patients affected by rheumatic diseases. Machine learning is a subfield of artificial intelligence that introduced a novel paradigm in programming methods. It can be defined as the ability of computers to learn how to solve a given problem without being explicitly programmed for this. In this work several supervised machine learning algorithms were employed to evaluate cardiovascular risk on rheumatic patients. Results of this explorative study open interesting perspectives for future developments of risk predictors.

Acknowledgements

First, I would like to sincerely thank my supervisors Prof. Marco Agostino Deriu for giving me the opportunity to work on such an amazing topic and for stimulating me to always push the boundaries of my knowledge, his passion and commitment for his work has greatly inspired me and Prof. Alberto Audenino for his guidance and fruitful advices through this master's degree journey.

Second, most of my gratitude goes to Eng. Giacomo Di Benedetto, for having found me as a totally inexpert student and having introduced me to the suggestive world of statistics and data science, with always brilliant hints and explanations, patience and gentleness. And to Lorenzo Pallante for his critical review of my work.

I would like to express my gratitude to Prof. Antonella Afeltra group, from Rheumatology Unit, Campus Bio-Medico University Hospital, Rome, to Dr. Luca Navarini and Dr. Domenico Paolo Emanuele Margiotta for having shared clinical data that I used in my work.

Then, the biggest thank is for people who have supported me in these long and though years, in many ways. To my mother and my father, who have always listened to me and without whom I would not be the person that I am today. This degree is dedicated to you. To my brother Luca, the best friend I could ever asked for. To my dear friend Alessandra, who taught me to believe in myself and in women's power. To my beautiful "BG" Chiara and Irene, our cocktails after exams will be my best memories of these years. To my colleagues Stefano, Alessandro, Federico, Fabio, Miriam, Miriana, Luca, Carlo, Chiara and Irene (an acquired colleague) for the great experiences shared together and for their true friendship. And finally, to the friends of a lifetime, from a small town but with a big heart.

1. Introduction

Cardiovascular (CV) diseases are the first cause of death worldwide. Many efforts have been done in the field of CV risk prediction, to early identify people at high risk and to treat them correctly. This is possible through the study of risk factors: characteristics of a person associated with an increased risk of developing a specific disease, such as a CV disease. Risk factors are very important in clinics, because they are causal and modifiable (a defined benefit should result from their modification). Many risk estimation systems exist, including in general predictors like age, sex, smoking status, blood lipids level and blood pressure. The best known and most widely used risk score globally is surely Framingham risk score.

In patients affected by inflammatory arthritis an increased CV risk has been observed and the performance of traditional CV risk predictors is a largely debated subject. CV complications deeply affect rheumatic patients' life, therefore finding the right treatment for each patient is of crucial importance. Traditional risk algorithms underestimate the risk in rheumatic patients. To overcome this limitation, the following actions have been proposed: the European League Against Rheumatism (EULAR) recommended to adapt general population risk algorithms with a multiplication by the factor of 1.5; also redefining cut-off values seems a reasonable strategy, having a similar effect as the corrective EULAR coefficient.

More specific CV risk algorithms are needed in the case of rheumatic patients, including novel biomarkers and disease-related CV risk predictors, together with prospective and larger studies. In this vision, the aim of this research is to explore traditional risk predictors performance on the general population and on the rheumatic one (specifically patients with psoriatic arthritis, ankylosing spondylitis and systemic lupus erythematosus), trying to evaluate and rationalize the use of cut-off strategy and EULAR correction coefficient. Then, we tried to apply the novel machine learning (ML) techniques to face CV risk prediction in rheumatic patients, comparing the results with the traditional ones, with the addition of a features analysis study. ML belongs to the broader field of artificial intelligence and was born with the idea of developing intelligent systems able to learn how to solve a specific problem without being explicitly programmed for it. They derive knowledge from big quantities of data. The three main subfields of ML are supervised learning, unsupervised learning and reinforcement learning. In this study, supervised learning was

adopted to predict the CV risk from a database of patients for which the final event was already known, but in the future also clustering techniques may be employed to discover new subgroups of the disease. This approach is innovative, in fact ML techniques have never been applied before to cardiovascular risk prediction in rheumatic patients.

This thesis is divided in the following sections:

Chapter 1 is the present introduction.

Chapter 2 is dedicated to the medical background behind this work, i.e. the concept of cardiovascular prediction, the description of the rheumatic pathologies we analyzed (psoriatic arthritis, ankylosing spondylitis and systemic lupus erythematosus) and the observed increase of CV risk in them.

Chapter 3 is an overview of the methods used in this thesis. ML methods are first generally described, through their typical workflow and all the stages required to perform classification. Then, a more precise description of the employed algorithms (i.e. support vector machine, random forest and k-nearest neighbors) is reported, together with the technique used for features analysis.

Chapter 4 describes the most various applications of ML inside biomedical engineering. ML have been applied in almost every sector, from imaging to signal analysis, from diagnosis to treatment aid. Also, the cutting-edge deep learning techniques have been explored. Finally, a specific paragraph has been dedicated to ML for cardiovascular risk prediction.

Chapter 5 is devoted to the investigation of ML techniques (e.g. random forest) application to the prediction of CV risk in general and rheumatic patients (with psoriatic arthritis, ankylosing spondylitis and systemic lupus erythematosus), comparing results with those of traditional risk prediction algorithms (e.g. Framingham risk score). Also, feature analysis has been performed to explore the importance of traditional predictors and novel rheumatic ones.

2. Background

In patients affected by **chronic inflammatory arthritis**, such as rheumatoid arthritis, psoriatic arthritis and axial spondyloarthritis an **increased cardiovascular (CV) risk** has been observed combined with major adverse CV events (Yim and Armstrong, 2017). The proposed mechanisms for shared pathogenesis between psoriatic disease and cardiovascular ones are inflammation, insulin resistance, dyslipidemia, angiogenesis, oxidative stress and endothelial dysfunction. There are complex relationships and intersections among these mechanisms that are not solved yet. While a lot is known about the epidemiology of psoriatic and CV diseases, the understanding of shared mechanisms is still missing. In Figure 1 a summary of the possible interconnections among mechanisms and clinical factors is represented.

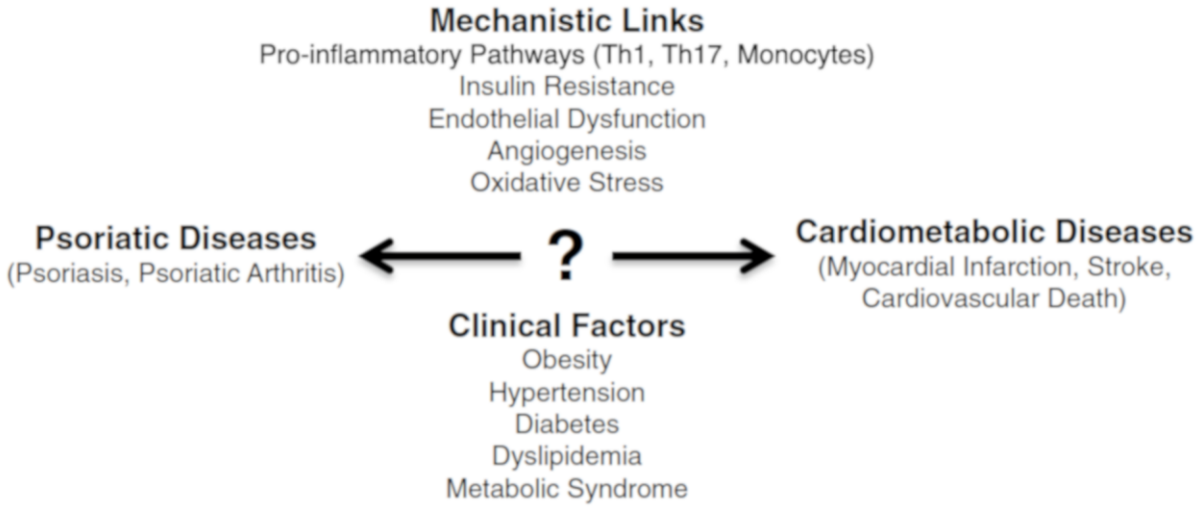

Figure 1 - Current proposed mechanisms and clinical factors that can cause the contribution of psoriatic diseases to cardiovascular diseases.

Future translational research is needed to investigate the link between psoriatic diseases and CV ones and to find new clinical ways to improve the lives of psoriasis patients. Moreover, identification of patients at high CV risk is mostly important to identify preventive strategies, like lifestyle changes and pharmacological interventions.

2.1 Psoriatic Arthritis, Ankylosing Spondylitis and Systemic Lupus Erythematosus

Psoriatic arthritis (PsA) is an inflammatory arthritis linked to personal or familiar history of psoriasis. It is characterized by a heterogeneous clinical manifestation and course, with possible axial and/or peripheral joint involvement, enthesitis and dactylitis (Ritchlin, Colbert and Gladman, 2017). It is classified inside the group of seronegative spondyloarthritis (SpA), which presents common laboratory, clinical and radiological features. There are different types of seronegative SpA, not necessarily representing distinct diseases, but overlapping significantly in etiology, pathology, clinical features and treatment:

- Ankylosing spondylitis (most common)
- Reactive arthritis
- Psoriatic arthritis
- Undifferentiated spondyloarthritis
- Spondyloarthritis associated with Crohn's disease and ulcerative colitis

The reasons behind the numerous uncertainties about this pathology are that for long time a unique clinical definition was missing. In 1964, the American rheumatism association has recognized for the first time PsA as a distinct articular inflammatory pathology.

PsA is common in Caucasian population (with a prevalence of 1-3%), while it is less present in other ethnic groups such as the Afro-Americans (with a prevalence of 0-0.3%). Disease distribution in male and female seems heterogeneous and the onset age is included between 30 and 50 years, but it can also appear during childhood. In most cases (67%), psoriasis onset comes first than arthritis; in 16% of cases, psoriasis and arthritis appear within 12 months from one another and in about 15% of cases, psoriasis can follow arthritis of some years. While data about the prevalence of psoriasis in the general population are certain, the percentage of psoriatic patients that develops rheumatic symptoms is still debated, because in literature have been reported very different values (between 0.2% and 42%). An Italian study (Salvarani *et al.*, 1995) on the prevalence of arthritis in 205 patients with psoriasis has highlighted the presence of arthritis in 36% of subjects. Currently, in Italy, the prevalence of PsA among psoriatic patients is believed to be between 7.7% and 35%. PsA etiology is not known, but a multifactorial origin, linked to genetic, environmental and immunological factors is hypothesized. Recent studies have demonstrated the central role of some

cytokines (such as TNF-α, IL-12, IL-23 and IL-17) in the determination of the inflammatory process and the structural damage.

PsA is a disease in which different clinical situations can coexist. PsA is commonly divided into 5 clinical groups (following Wright and Moll classification, 1973): asymmetric oligoarthritis (60-70% of cases), the form involving distal interphalangeal joints of hands and feet (5-10% of cases), the mutilans form (1-2% of cases), the symmetric polyarthritis (15-20% of cases) and spondylitis (5-10% of cases). Clinically, PsA does not differ a lot from rheumatoid arthritis: patients report pain and prolonged morning rigidity, with warm and swollen involved joints. Typical manifestations are: dactylitis (Salvarani, Gabriel and Hunder, 1996), the so-called sausage digit (see Figure 2), which can become chronic and poorly responsive to therapy; enthesitis (D'Agostino, Palazzi and Olivieri, no date), that is enthesis inflammation (enthesis is the connective tissue between tendon or ligament and bone), and is also present in other kinds of spondyloarthritis; onychopathy, which is characterized by nails distal thickening and other associated symptoms such as uveitis.

Figure 2 - Dactylitis: evidence of severe swelling of the third finger of the left hand.

Objective and functional measurements are very useful to study and monitor the disease course (Mease, 2005). Count of the number of involved joints with swelling and/or pain is one of the most used parameters to evaluate the disease. An index already employed for rheumatoid arthritis is applied also to PsA: the 28 joints disease activity score (DAS28). Another useful index is PASI (psoriasis area and severity index), with the aim of evaluate psoriasis severity. The body is divided

into four parts: head, torso, superior and inferior limbs. For each section, psoriasis area and seriousness are verified, with respect to three parameters: erythema, infiltration and desquamation. Each of them can be judged as absent, light, moderate, sever and very severe. The total score is obtained as the sum of the three partial scores.

The ideal therapy should control every disease aspect: the joint, axial and cutaneous ones. Non-steroidal anti-inflammatory drugs, cortisone and disease-modifying antirheumatic drugs are widely employed in PsA treatment.

Ankylosing spondylitis (AS) is an autoimmune, inflammatory disease that mainly involves the axial skeleton and the sacroiliac joints. It is prevalent in the western world with an estimated incidence of 7/100000 per year (0.1-0.2% if the population is affected by this pathology). It presents chronic characteristics which make it an affecting physical function and quality of life pathology. Comorbidities such as uveitis, inflammatory bowel disease, psoriasis, aortic insufficiency and osteoporosis contribute the worse the pathology and the prognosis (Wang and Ward, 2018). Recently, both mortality and morbidity ratios of CV diseases have been found to be increased in people with AS. It is known that patients affected by AS are at increased risk of CV diseases (Kinsella, Johnson and Ian, 1974). This may be a consequence of the typical chronic inflammation of AS that manifests itself in cardiac structures (Bengtsson *et al.*, 2017); the role of inflammation in atherogenesis and plaque formation is now established and inflammation has been connected to an atherogenic lipid profile among people with AS. Other factors may include the use of nonsteroidal anti-inflammatory drugs, decreased physical activity, genetics and the higher frequency of metabolic syndrome among this kind of patients.

Systemic lupus erythematosus (SLE) is a chronic autoimmune pathology of the connective tissue, characterized by a multifactorial pathogenesis with a strong interconnection between genetic and environmental factors, which determines the disease course (Tsokos, 2011). The production of a wide range of autoantibodies is a typical characteristic of this pathology, that leads to different clinical phenotypes (Yaniv *et al.*, 2015). SLE pathophysiological mechanism is complex and it is characterized by the immune system disfunction, promoted by genetic, epigenetic and environmental factors, which causes the tissue damage. SLE has the capability to transversally affect every organ and tissue of the organism, with different intensity and in various ways during pathology course. In fact, the typical symptoms are divided into: constitutional symptoms, like

fatigue (with depression, sleep disorders, smoking habit, sedentary life…), fever and weight changes, frequent at the beginning of the disease and easily misunderstood, with the consequent diagnostic delay; organ symptoms and signs, among which musculoskeletal and mucocutaneous symptoms, renal, pulmonary, gastrointestinal, cardiovascular and eye involvement, neurological and vascular manifestations and finally hematological alteration. The pathology course alternates disease activity phases with remission phases. With the increasement of SLE patients' survival, a progressive reduction of the activity with age has been noticed, but on the other hand new manifestations have also been seen, due to the prolonged inflammatory state and the chronic drugs assumption (Sutton, Davidson and Bruce, 2013).

Typically, it shows its first manifestations between the age of 16 and 55 years, with prevalent incidence in female (9:1). However, in early or late onsets, the incidence is similar in both male and female, with a slightly prevalence in male in the case of early lupus. Moreover, the disease prevalence depends on patients' ethnicity, from 120/200 cases on 100000 in Afro-Americans and Caribbean population (Hopkinson, Doherty and Powell, 1994) to 50 cases on 100000 in Asians population, until 12.5 on 100000 in Caucasian population.

SLE prognosis is of 95% at 5 years, of 90% at 10 years and 78% at 20 years. In the last 50 years, SLE prognosis has increased (Doria *et al.*, 2006), but at the same time new comorbidities appeared, which determined a radical change of mortality causes in SLE patients. For example, in the chronic phase of the pathology, cardiovascular damage, tumors and drugs damage determine patients' death. According to these evidences, the identification of new methods able to predict the accrual and the progression of SLE damage is a strategic objective with the final aim of identifying patients at high risk.

In order to quantify the damage in SLE patients and to measure how the disease changes over time, the systemic lupus collaborating clinics (SLICC) and the American college of rheumatology (ACR) proposed and validated an index, the SLICC/ACR damage index, SDI (Gladman *et al.*, 1996). It consists of 12 different organ systems. Table 1 shows how to assign scores to patients on the base of organs involvement.

Table 1 - SLE SLICC/ACR damage index (SDI) table to assign scores to patients.

Item	Score
Ocular (either eye, by clinical assessment)	
Any cataract ever	1
Retinal change *or* optic atrophy	1
Neuropsychiatric	
Cognitive impairment (e.g., memory deficit, difficulty with calculation, poor concentration, difficulty in spoken or written language, impaired performance level) *or* major psychosis	1
Seizures requiring therapy for 6 months	1
Cerebrovascular accident ever (score 2 if >1)	1 (2)
Cranial or peripheral neuropathy (excluding optic)	1
Transverse myelitis	1
Renal	
Estimated or measured glomerular filtration rate <50%	1
Proteinuria ≥3.5 gm/24 hours	1
or	
End-stage renal disease (regardless of dialysis or transplantation)	3
Pulmonary	
Pulmonary hypertension (right ventricular prominence, or loud P2)	1
Pulmonary fibrosis (physical and radiograph)	1
Shrinking lung (radiograph)	1
Pleural fibrosis (radiograph)	1
Pulmonary infarction (radiograph)	1
Cardiovascular	
Angina *or* coronary artery bypass	1
Myocardial infarction ever (score 2 if >1)	1 (2)
Cardiomyopathy (ventricular dysfunction)	1
Valvular disease (diastolic, murmur, or systolic murmur >3/6)	1
Pericarditis for 6 months, *or* pericardiectomy	1
Peripheral vascular	
Claudication for 6 months	1
Minor tissue loss (pulp space)	1
Significant tissue loss ever (e.g., loss of digit or limb) (score 2 if >1 site)	1 (2)
Venous thrombosis with swelling, ulceration, *or* venous stasis	1
Gastrointestinal	
Infarction or resection of bowel below duodenum, spleen, liver, or gall bladder ever, for cause any (score 2 if >1 site)	1 (2)
Mesenteric insufficiency	1
Chronic peritonitis	1
Stricture *or* upper gastrointestinal tract surgery ever	1
Musculoskeletal	
Muscle atrophy or weakness	1
Deforming or erosive arthritis (including reducible deformities, excluding avascular necrosis)	1
Osteoporosis with fracture or vertebral collapse (excluding avascular necrosis)	1
Avascular necrosis (score 2 if >1)	1 (2)
Osteomyelitis	1
Skin	
Scarring chronic alopecia	1
Extensive scarring or panniculum other than scalp and pulp space	1
Skin ulceration (excluding thrombosis) for >6 months	1
Premature gonadal failure	1
Diabetes (regardless of treatment)	1
Malignancy (exclude dysplasia) (score 2 if >1 site)	1 (2)

Typical therapies are based on corticosteroids, antimalarial drugs, immunosuppressants and biological drugs.

2.2 Cardiovascular Risk and Rheumatic Diseases

In many inflammatory systemic pathologies (such as the already mentioned PsA, AS and SLE) an increase of CV risk has been reported, which starts with a subclinical atherosclerosis. Atherosclerosis is a chronic vascular pathology, characterized by the inflammation of big and medium size arteries' intima (the more internal layer, the one in direct contact with blood). Inflammation is mostly caused by chronic lipid-driven inflammatory disease, which produces many dynamic pathological situations, among which the most typical is the atherosclerotic plaques formation. Atherosclerosis is characterized by a long preclinical phase, which happens between the beginning of the atherogenic process and the clinical disease manifestations (among which stroke and acute myocardial infarction). In this asymptomatic phase there are functional and morphological vascular modifications, both explorable (for example by means of echography and color Doppler).

CV disease development in rheumatic pathologies involves genetic factors, modifiable risk factors (dyslipidemia, diabetes mellitus, hypertension), inflammatory components of the immune response and the autoimmune elements, like autoantibodies, autoantigens and autoreactive lymphocytes (McMahon and Hahn, 2007). Indeed, epidemiological and cohort studies have demonstrated that in the atherosclerosis pathogenesis not only traditional risk factors are involved, but also inflammation with its mediators, like cytokines, chemokines, T and B lymphocytes and antibodies (Ross, 1999; Shoenfeld, Sherer and Harats, 2001). In atherosclerotic plaques there can be found abundant infiltrated of T lymphocytes and Th1 are predominant with respect to Th2. Recently it has been demonstrated that CRP, in SLE patients, forms immunocomplexes which circulate and deposit in arteria intima. A systematic review conducted on 14 studies made between 1980 and 2009 (Roifman *et al.*, 2011) demonstrated that patients affected by chronic inflammatory conditions are at high risk of developing a coronary heart disease. This risk does not come only from the most common systemic inflammatory diseases (like rheumatoid arthritis and SLE), but also from less common pathologies, such as AS, dermatomyositis and PsA. Thus, the link between systemic inflammatory pathologies and CV diseases can be only partially explained by means of

traditional CV risk factors, which present a high prevalence in rheumatic patients. It is hypothesized that a prominent role in this link could be played by the inflammatory immune-mediated pattern, activated in rheumatic diseases, and by potential adverse effects caused by corticosteroids and anti-inflammatory non-corticosteroids. CV involvement in rheumatic diseases is extremely heterogeneous, with effects on pericardium, cardiac muscle, endocardium and valves, conduction system, coronary arteries, systemic circulation big and small vessels, with severe consequences such as the augmented risk for ischemic cardiopathy and stroke in earlier age. Table 2 summarizes the principle effects induced by rheumatic diseases on CV system.

Table 2 - CV involvement in the principal systemic inflammatory pathologies.

Rheumatoid arthritis	Heart attack
	Stroke
Systemic lupus erythematosus	Heart attack
	Stroke
	Pericarditis
	Myocarditis
	Conduction system anomalies
Systemic sclerosis	Myocarditis
	Pericarditis
	Heart failure
	Ventricular arrhythmia
	Atherosclerosis
Psoriasis	Heart attack
	Thrombophlebitis
	Pulmonary embolism
	Stroke

A better control of the inflammatory process with drugs that are also cardioprotective, an early rehabilitation, a precise control of the lipid, pressor and metabolic profiles, associated to targeted and effective screening strategies, can contribute to reduce the impact of the most common systemic inflammatory pathologies on CV system.

2.3 CV Risk Prediction in the General Population: Framingham Risk Score, CUORE Risk Score, SCORE Risk Score

The identification of people at high CV risk is one of primary individual prevention main objectives and it is the necessary premise to activation of actions aimed at reducing modifiable risk factors,

changing lifestyle and using pharmacological treatments. General population is here intended as generic people who present some of traditional CV risk factors at a level that makes them at potential risk of developing CV diseases. At the end of the 1980s, primary prevention guidelines were based on treatment of single risk factors. Recently, the attention has been drawn to the absolute global CV risk, indicator of the disease incidence, predictable on the base of the principal risk factors levels (Damen *et al.*, 2016). To identify subjects who present a high probability of having the disease, risk functions are used, derived from longitudinal studies conducted on population groups followed over time. Risk function appropriateness depends on study population characteristics and on subjects to whom they are applied. This means that when applying to an Italian population a risk function derived from an American population, like still happens, can create risk estimate distortions.

The main risk factors are divided into modifiable and not modifiable. Lifestyle plays an important role too.

Not modifiable risk factors:

- Age
- Gender
- Familiar history

Modifiable risk factors:

- Hypertension
- High level of LDL cholesterol
- Low level of HDL cholesterol
- High level of triglycerides
- Diabetes
- Obesity
- Thrombogenic factors

Lifestyle:

- Hypercaloric diet
- Smoking habit

- Alcohol abuse
- Sedentary life

There is an enormous quantity of models predicting incident CV disease in the general population. The median number of risk factors (predictors) included in the developed models is about 7. Figure 3 reports a histogram with the number of models including specific predictors.

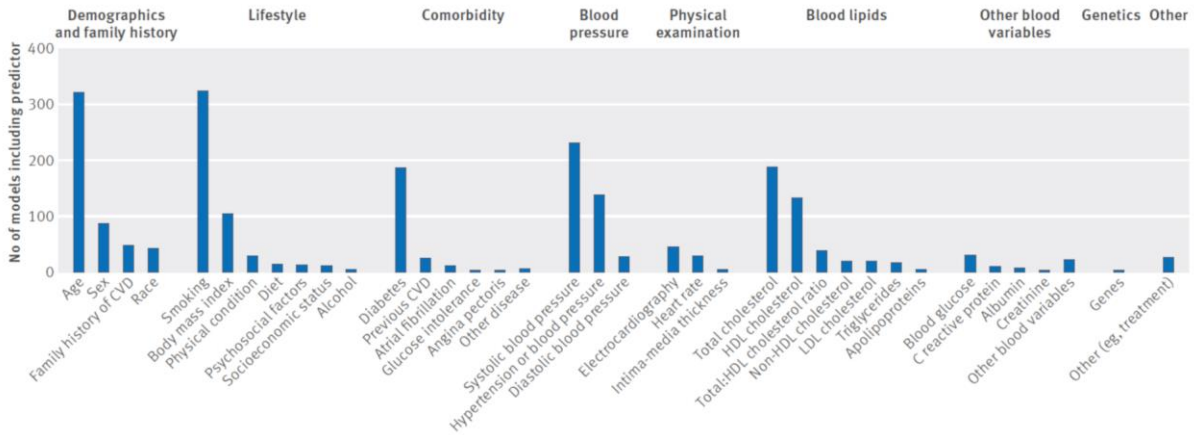

Figure 3 - Main categories of predictors included in developed models. CVD=cardiovascular disease, HDL=high density lipoprotein, LDL=low density lipoprotein.

The most common predictors are smoking habit and age and most models are sex specific. There is much variability in geographical location of models, but most of them were developed and validated in European and Northern American populations. A prediction model for people from Africa or South America has only recently been developed (Hajifathalian *et al.*, 2015).

Framingham risk score (FRS) is a single multivariable risk function that predicts the risk of developing all CV diseases, coronary, cerebrovascular, peripheral arterial disease and heart failure (D'Agostino *et al.*, 2008). It uses Cox-proportional hazards regression (Cox, 1972) to relate risk factors to the incidence of a first CV disease during a maximum follow-up period of 12 years. It was developed on 8491 participants (mean age, 49 years, 4522 women) who attended a routine examination between 30 and 74 years of age and were free of CV diseases. Given the predominantly white Framingham sample, the transportability of the model to other samples must be carefully evaluated. The algorithm demonstrated good discrimination (C statistic of 0.763 for men and of 0.793 for women) and calibration.

It is widely recognized that age, sex, high blood pressure, smoking, dyslipidemia and diabetes are the major risk factors for developing CV diseases. Moreover, CV disease factors are believed to cluster and interact multiplicatively (Jackson *et al.*, 2005). Therefore, covariates included in Cox models are age, total cholesterol, HDL cholesterol, systolic blood pressure (SBP), antihypertensive medication use, current smoking and diabetes status. All the continuous factors have been naturally logarithmically transformed to enhance the discrimination and calibration of the models and to minimize the presence of extreme observations. CV disease is defined as a composite of CHD (coronary death, myocardial infarction, coronary insufficiency and angina), cerebrovascular events (including ischemic stroke, hemorrhagic stroke and transient ischemic attack), peripheral artery disease (intermittent claudication) and heart failure.

The model's equation is the following:

$$FRS = 1 - S_0(t)^{\exp(\sum_{i=1}^{p} \beta_i X_i - \sum_{i=1}^{p} \beta_i \bar{X}_i)}$$

where $S_0(t)$ is the baseline survival at follow-up time t (here t = 10 years), β_i the estimated regression coefficient (log hazard ratio), X_i the log-transformed value of the i-th risk factor (if continuous), $\bar{X}_i$ the corresponding mean and p the number of risk factors. The multivariable-adjusted regression coefficients and hazard ratios for CVD prediction are presented in Table 3. The value obtained from the model's equation corresponds to a risk percentage, estimated at 10 years.

Table 3 - Regression coefficients and hazard ratios. $S_0(10)$ indicates 10-year baseline survival and β the estimated regression coefficients.

Variable	**β women** **$S_0(10) = 0.95012$**	**Hazard ratio women**	**β men** **$S_0(10) = 0.88936$**	**Hazard ratio men**
Log of age	2.32888	10.27	3.06117	21.35
Log of total cholesterol	1.20904	3.35	1.12370	3.08
Log of HDL cholesterol	-0.70833	0.49	-0.93263	0.39

Log of SBP if not treated	2.76157	15.82	1.93303	6.91
Log of SBP if treated	2.82263	16.82	1.99881	7.38
Smoking status	0.52873	1.70	0.65451	1.92
Diabetes	0.69154	2.00	0.57367	1.78

FRS limitations are the poor applicability to the European population, especially the South-European one. In fact, FRS overestimates CV risk in the Italian population, because it has been developed for a North-American population with hypertension and with higher risk factor than those present in Europe (Bastuji-Garin *et al.*, 2002). Moreover, FRS does not account for some important CV risk factors, such as CV disease family history and weight; also, lifestyle (active or sedentary), which is important to evaluate the risk, is not considered.

CUORE risk score was built within the Italian CUORE project (for epidemiology and prevention of ischemic heart diseases). CUORE project was launched in 1998; it is financed by 1% of the national health fund and it is coordinated by the Istituto Superiore di Sanità. It is a prospective fixed-cohort study, which includes cohorts from the north, the center and the south of Italy. Since 2005, the project is included among those of the National Centre for Disease Prevention and Control, Ministry of Health, Rome (Doukaki, Caputo and Bongiorno, 2013). Figure 4 shows the available online survey to calculate individual risk. This score indicates how many people out of 100 of the same age, sex and characteristics will present a first CV event in the next 10 years.

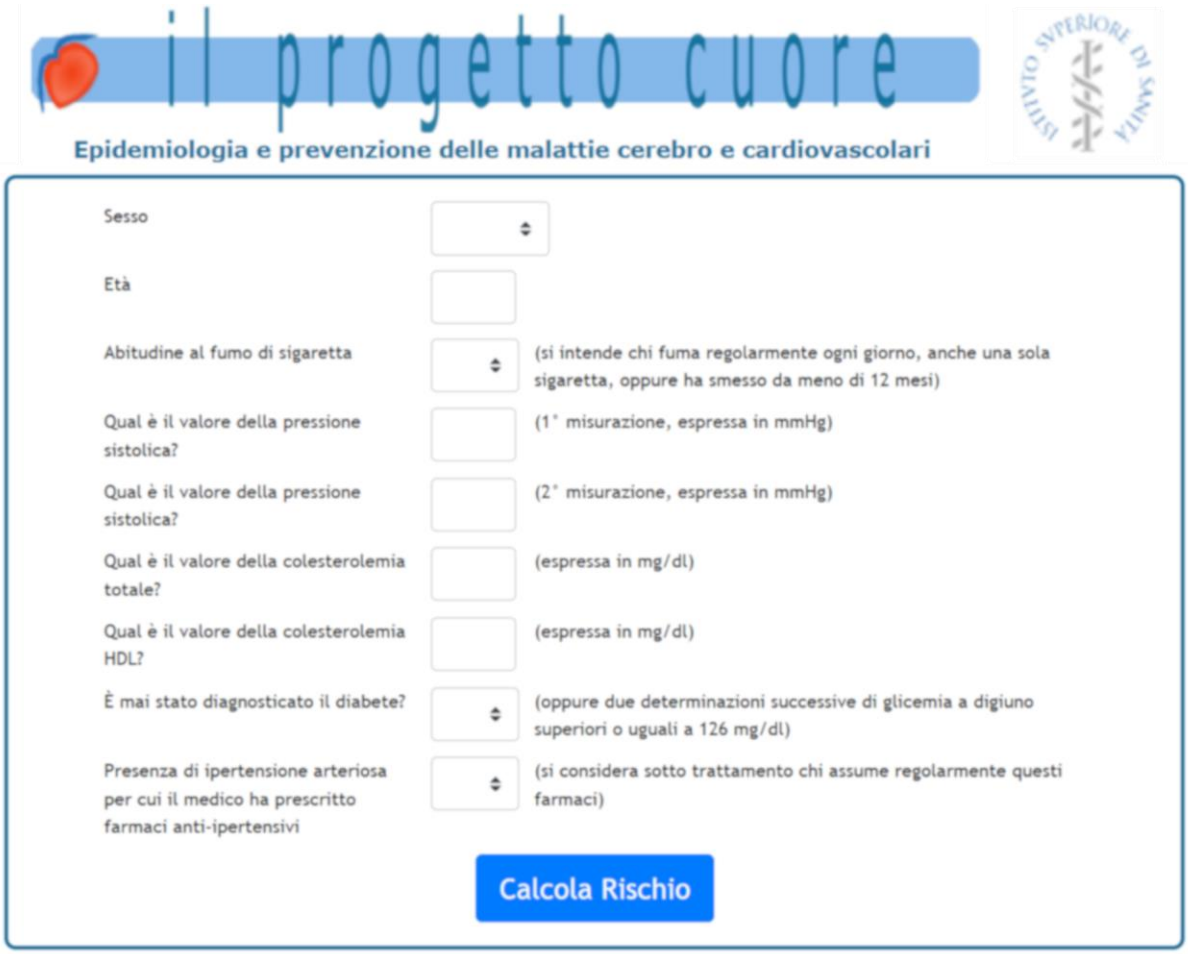

Figure 4 - Online survey for CV risk individual score calculation, inside Italian CUORE project (www.cuore.iss.it).

CUORE risk score allows evaluating the probability of experiencing a first CV event (myocardial infarction, stroke) over the next 10 years by a subject. It is easy to be applied to achieve a fast and objective estimation of the absolute global CV risk in primary prevention. The risk factors included in the model are: age, gender, total cholesterol, HDL cholesterol, SBP, hypertension treatment, smoking status and diabetes. The first major coronary or cerebrovascular event has been considered as endpoint. Out of 20647 people aged 35-69 years with no previous CV events, 971 major CV events (636 coronary and 335 cerebrovascular) have been identified. The assessment of risk factors coefficients has been performed by means of Cox proportional hazards model, separately for women and men.

The model's equation is the following:

$$CUORE = 1 - S_0(t)^{\exp(\sum_{i=1}^{p} \beta_i X_i - \sum_{i=1}^{p} \beta_i \bar{X}_i)}$$

where $S_0(t)$ is the baseline survival at follow-up time t (here t = 10 years), β_i the estimated regression coefficient, X_i the i-th risk factor, $\bar{X}_i$ the corresponding mean and p the number of risk factors. The regression coefficients for CVD prediction are presented in Table 4. The value obtained from the model's equation corresponds to a risk percentage, estimated at 10 years.

Table 4 - Risk factors β coefficients and S(t) survival which determinates CV risk function at 10 years separately for women and men.

Variable	**β women** **$S_0(10) = 0.989$**	**β men** **$S_0(10) = 0.953$**
Age	0.079	0.076
Total cholesterol	0.003	0.006
HDL cholesterol	-0.015	-0.013
SBP	0.016	0.013
Hypertension treatment	0.590	0.490
Smoking status	0.773	0.508
Diabetes	0.339	0.462

With respect to FRS, CUORE uses the same risk formula (derived by Cox regression), but different risk factors and different regression coefficients. Risk factors are almost the same, but in CUORE they are not logarithmically transformed, and hypertension treatment is considered as a binary feature, while in FRS only SBP (continuous value) associated to treated patients or not is considered. Regression coefficients are different, because while FRS has been developed on an American population, CUORE has been developed on an Italian one. Therefore, when these risk scores are transferred to a different population from that on which they were developed, it is convenient to perform a calibration (i.e. mean values in the risk formula are calculated based on the current population risk factors).

Since population risk changes over time, because it depends on the risk factors means over the population and by population survival without the disease, the systems to evaluate risk must be updated and reflecting the current lifestyle. Therefore, to use CUORE in the future, it is necessary to recruit new more recent cohorts.

Systematic coronary risk evaluation (SCORE) was developed inside a European project, because it was impossible to apply FRS to the different European populations, according to its authors (Conroy, 2003). This big researchers' group, belonging to 12 different European nations, believed that it was necessary to calculate an exclusively European score and differentiate it between high CV risk countries and low CV risk countries.

SCORE estimates the 10-year risk of developing a first fatal CV event. CV disease is defined as: heart attack, stroke, aneurysm of the aorta or other. Risk charts were built, using constant variables such as gender, age, SBP, smoking status and total cholesterol, without distinguishing between diabetic patients or non-diabetic ones. These charts were built for high risk countries (such as Denmark, Great Britain and Norway) and for low risk countries (such as Italy, Spain and Belgium). The advantage of this score is its simplicity, in fact it is not based on a mathematical formula, but on a score system. Moreover, this study has the merit to have considered absolute mortality due to CV events and not only to coronary events as outcome. Figure 5 shows an example of SCORE chart. SCORE gives the possibility to calculate the estimated risk for each patient with an indication of the age as number of exposure years to the risk. This concept overcomes one of biggest limits of CV scores: older patients with the same factors as younger patients have a higher risk. This is only partially true: a younger patient can have more exposure years to factors that influence his or her prognosis, therefore, more than age, exposure years influence the CV risk.

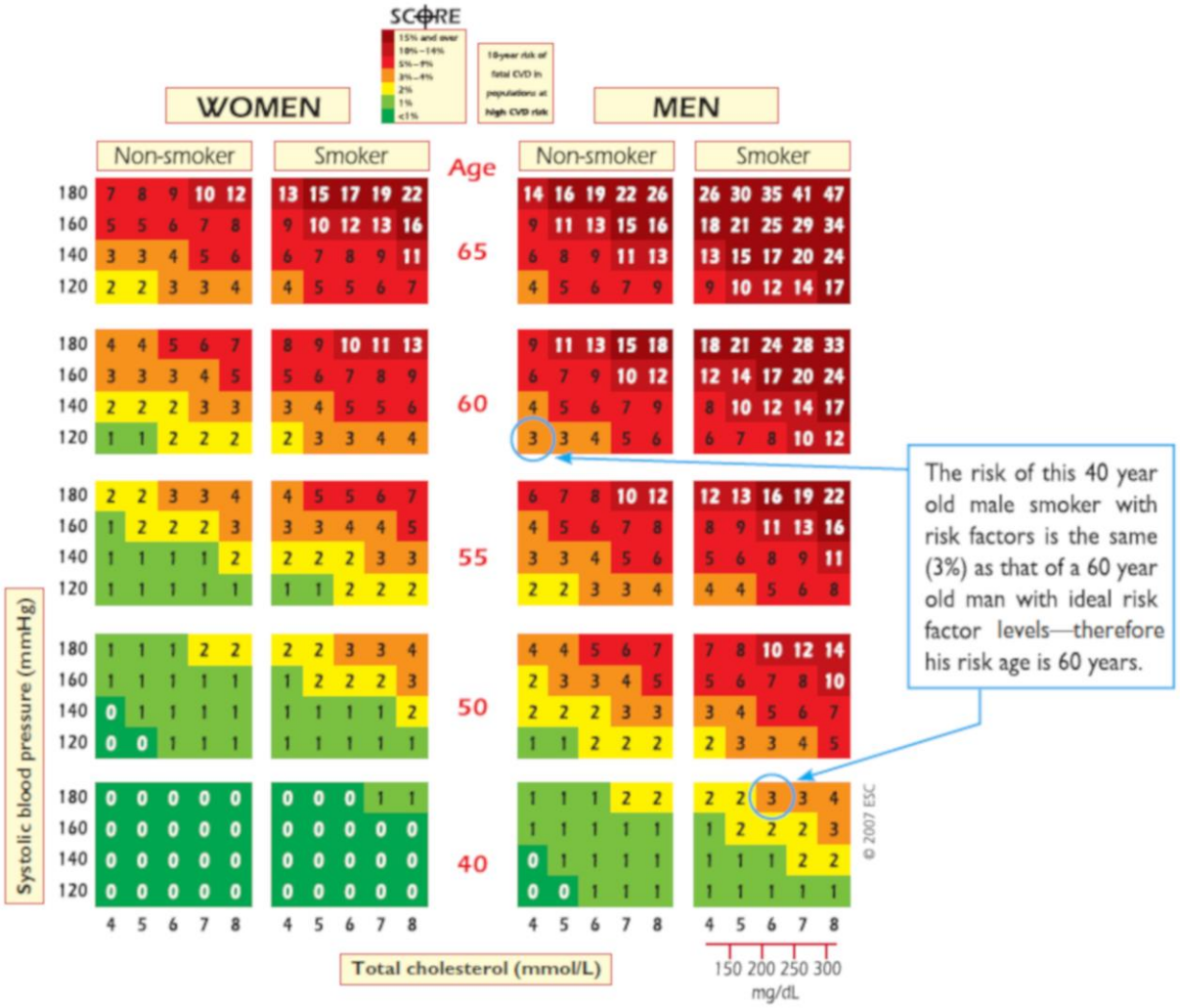

Figure 5 - An example of SCORE chart with an illustration of risk-age concept.

Table 5 summarizes and compares the different risk factors included in each of the explained algorithms.

Table 5 - Items included in SCORE, CUORE and FRS.

Baseline characteristics	Framingham	CUORE	SCORE
Age (years)	✓	✓	✓
Gender (male/female)	✓	✓	✓
Systolic blood pressure (mmHg)	✓	✓	✓
Total cholesterol (mg/dl)	✓	✓	✓
HDL cholesterol (mg/dl)	✓	✓	
Smoking status (Y/N)	✓	✓	✓
Diabetes mellitus (Y/N)		✓	
Use of antihypertensive medication (Y/N)	✓	✓	

2.4 Cardiovascular Risk Prediction in Rheumatic Patients

The identification of high CV risk patients with rheumatic diseases such as PsA is very important, to identify preventive strategies. The performance of general CV risk algorithms in patients with inflammatory arthritis is a largely debated subject. Arts and coworkers (Arts *et al.*, 2015) assessed the performances of four CV risk algorithms in evaluating the risk of fatal and non-fatal CV events in European patients with rheumatoid arthritis. The results showed that all algorithms (FRS and SCORE were included in this study) tend to underestimate CV risk in patients with rheumatoid arthritis meaning that the real risk exceeds that predicted. Moreover the different scores appear poorly calibrated for these patients (Crowson *et al.*, 2012). Ernst and coworkers demonstrated that FRS underestimates CV risk in patients with PsA (Ernste *et al.*, 2015). Recently, the European league against rheumatism (EULAR) recommended to adapt the general population risk algorithms by means of a multiplicative factor of 1.5 in patients with inflammatory arthritis.

In a recent study, Navarini and coauthors (Navarini *et al.*, 2018) evaluated the performance of five original and adapted according to EULAR recommendations CV risk algorithms in PsA: SCORE, CUORE and FRS were analyzed in the study. They used prospectively collected data from two Italian cohorts and calculated sensitivity and specificity in the cases of low-to-intermediate (10%, except for SCORE: 1%) and intermediate-to-high (20%, except for SCORE: 5%) risk cutoffs. These cutoffs are thresholds put on the probability risk scores; their values are taken from literature and they differentiate patients with CV event from patients without, according to the algorithm's result. More than one cut-off is used, because they have different classification performances in terms of sensitivity and specificity. Results showed that discriminative ability and calibration were not improved by adaptation of the algorithms according to EULAR recommendations. Overall, the five scores evaluated in Navarini's study underestimated the risk. Although in PsA an increased CV risk has been observed, specific CV risk algorithms for this type of patients do not exist. Consequently, it seems crucial to redefine cut-off values for low and intermediate CV risk in patients with PsA, because underestimation of CV risk in patients with PsA could lead to insufficient treatment. Other possibilities are adding more biomarkers and disease-related CV risk factors in prediction models to provide PsA-specific CV risk scores.

3. Machine Learning Techniques

Machine learning (ML) is a subfield of *artificial intelligence* (AI) that can be defined as the ability of computers to learn how to solve a given problem without being explicitly programmed for this. Figure 6 shows the relationships between all the different fields present in *data science*, to whom machine learning belongs.

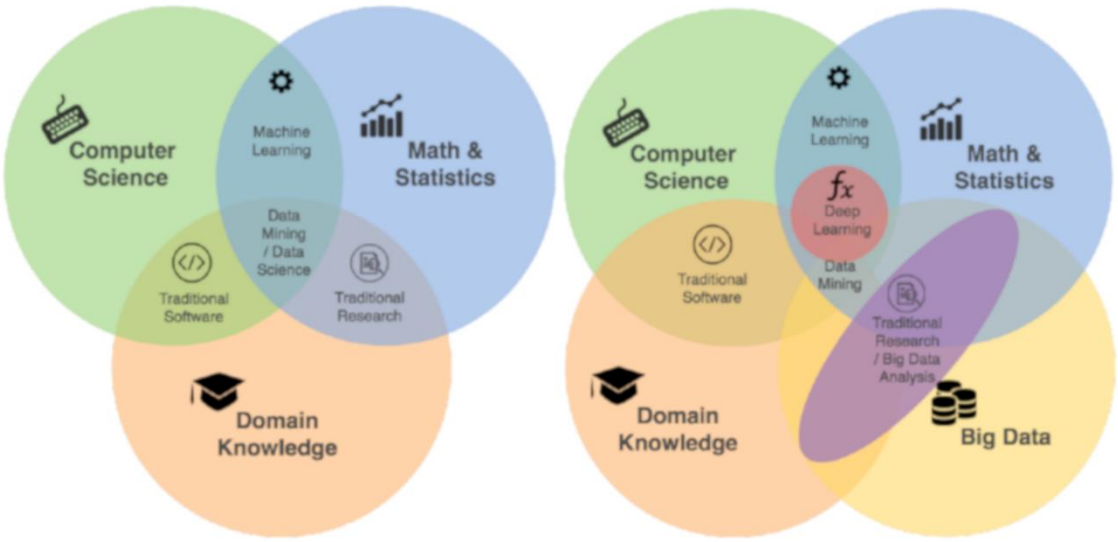

Figure 6 - Venn diagram of data science.

The term was first coined by Arthur Samuel (Samuel, 1959). The learning process is possible by deriving knowledge from experimental data and has the objective of making predictions. Nowadays, data is a resource present in huge quantity in almost every field. ML is a change of paradigm from knowledge-based algorithms to learning systems.

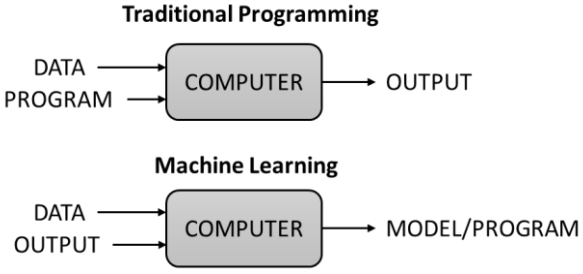

Figure 7 - The change of paradigm from knowledge-based systems (which use inference) to learning systems (which use induction).

Figure 7 shows that knowledge-based algorithms are a top-down approach, because they create a model of already acquired knowledge in a program that performs a specific task, while machine learning systems are based on a bottom-up approach, because they extract knowledge from examples to produce a final model that can continuously self-improve and solve problems. These kinds of methods are necessary when human expertise lacks, when it is not explainable, unreliable or unfeasible and when solutions may change over time or need to be adapted to specific cases. ML contains an ensemble of techniques that can be exploited inside *data mining* (DM). DM is defined as the process of automatically discovering patterns in data, which is stored electronically in databases. There are three big types of machine learning techniques: **supervised learning**, **unsupervised learning** and **reinforcement learning**. Supervised learning algorithms contain previous knowledge about data (i.e. labels describing the desired output of the model) and need to be trained using this knowledge before being applied to completely new data. Two types of supervised learning exist: **classification** and **regression**. Unsupervised learning algorithms deal with data of unknown structure (i.e. without labels) and are potentially able to discover new associations between inputs. Typical applications are **clustering** and **dimensionality reduction**. In reinforcement learning, instead, the goal is to develop a system that can improve its performances through its interactions with the environment (quantified by a reward, a measure of how good the action taken by the system was). The main applications area are game theory and robotics.

The main goal of supervised learning is to build a model from a dataset (that already contains desired outputs) and to use it to make predictions on future data or data for which desired outputs are not present (see Figure 8).

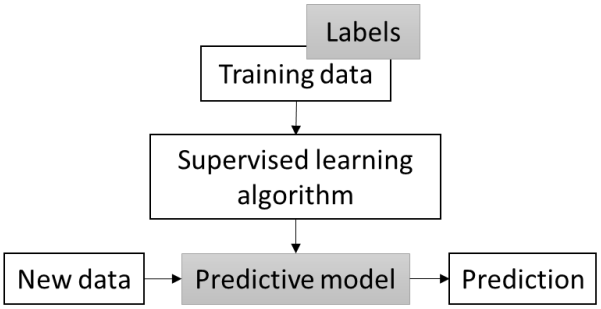

Figure 8 - Supervised learning approach workflow. Labels associated to data are used to build a predictive model to be employed to new unknown data.

Classification and regression are two subcategories of supervised learning. Classification has the goal of predicting categorical class labels, while regression predicts continuous outcomes; the first is used when dealing with discrete labels and outputs, the second when dealing with continuous ones. The simplest form of classification is the binary one: the algorithm learns how to discriminate two possible classes, for example the healthy class and the ill class. The word regression was first coined by Francis Galton in 1886 (Galton, 1886).

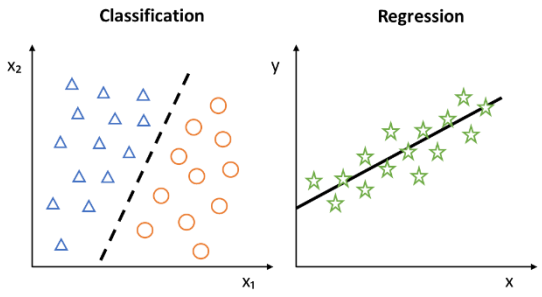

Figure 9 - The difference between classification (a binary example) and regression.

Figure 9 shows graphically the differences between classification and regression. In the case of classification, a binary example is represented, with a bidimensional dataset (each sample is characterized by two features, x_1 and x_2). There are two classes to which each object can be assigned (triangle class and circle class), by means of a decision function (represented by the dotted line) learnt by the supervised learning algorithm. In the case of regression, a continuous output must be predicted. There are some descriptive variables and a continuous target variable and the aim is to find a relationship between these two entities. In the figure, x is the predictive variable, while y the result and the straight line represents the model (calculating minimizing the distance between the points, here represented as stars) that can be used to predict the target of new data. Figure 10 represents the traditional machine learning pipeline, composed by four phases: preprocessing, learning, validation and prediction.

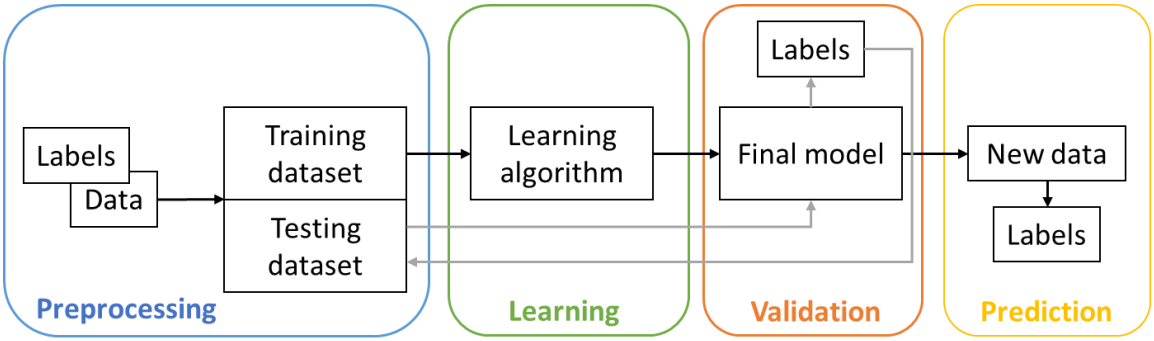

Figure 10 - A typical workflow to employ machine learning for predictive modeling.

Preprocessing is a fundamental step, because raw data rarely have a form that gives the learning algorithm optimal performances. On the contrary, they need to be transformed, for example extracting relevant features from them and performing features transformations so that they adopt the same scale. Data are usually divided into two different sets: the training set and the test set. The first one is used to train and optimize the algorithm, while the second is used to evaluate the final model and see if it has generalization capability. About learning and validation phases, a lot of different machine learning algorithms exist, with the aim of solving different problems. There is not an optimal algorithm for every problem, as the famous *No Free Lunch Theorems* states (Wolpert, 1996; Wolpert and Macready, 1997). Every algorithm has its pros and cons and without knowledge about the task to perform we cannot choose one algorithm as the best a priori. Therefore, typically, a certain number of algorithms are compared during the learning and validation phases, from different machine learning areas to finally select the model that offers the best performances. For this reason, the training dataset is divided into other validation subsets, to compare different models on a dataset that is different from the training one. The test set (which contains data that the model has never seen before) is then used to estimate the generalization performances of the model. In these two phases hyperparameters optimization techniques are also performed, to tune and adapt the model's parameters to the problem. Hyperparameters are model's coefficients that are not learnt form data, but that can improve model's performances. Finally, when the best model has been chosen, it can be used also to predict new future data.

3.1 An Introduction to the Classification Problem

Given a classification problem, the first questions to answer are:

- How should the objects to be classified be represented?
- What algorithm can be used?
- How should training be performed?
- How can classification performance be evaluated?

An **object** (sample, instance or observation) to be classified is represented by a set of **features** (parameters, attributes, dimensions or measurements) and this is generally depicted by a numerical vector. Therefore, the dataset is typically represented by a matrix, as Figure 11 shows. This is just a convention and each application require preprocessing to extract relevant features and represent features in the most convenient way to build an effective machine learning system, because real world data does not arrive in neatly aligned feature vectors.

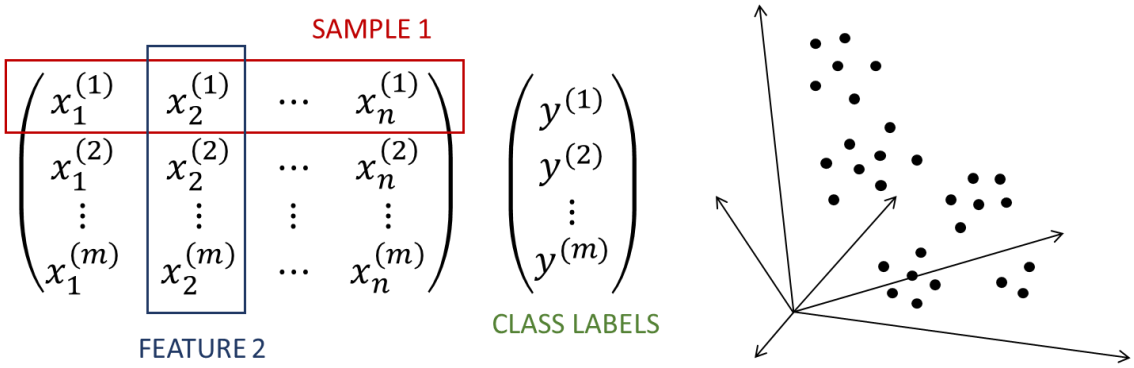

Figure 11 - On the left, the traditional representation of a dataset by means of a matrix. Instances are contained in the rows, while features are contained in the columns. An example of class vector is also shown. On the right, the feature space is represented.

Every sample can be imagined as a point in a n-dimensional **feature space**, where n is the number of features (Figure 11). Features may not be coherent with respect to each other (they can have different measure units, they can be numerical or not, they can be ordered or not, they can be discrete or continuous and they can have different ranges), therefore it is convenient to rescale the feature vector, so that no one of them can prevail over the others. Two approaches exist to transform different features over the same scale: **normalization** and **standardization**. Normalization means taking features on a scale from 0 to 1 and it is a special case of min-max scaling. It is applied to every column of the dataset and each new feature value is calculated as follows:

$$x_{norm}^{(i)} = \frac{x^{(i)} - x_{min}}{x_{max} - x_{min}}$$

where $x^{(i)}$ is a sample, while x_{min} and x_{max} are respectively the lowest and the highest number in the feature column. Standardization can be better for many machine learning algorithms, for example support vector machine, which initializes weights to 0 or small number near it. Indeed, standardization centers each feature to 0 mean and 1 standard deviation giving it a normal distribution, for which it is simpler to derive weights. Standardization is calculated as follows:

$$x_{std}^{(i)} = \frac{x^{(i)} - \mu_x}{\sigma_x}$$

where μ_x is the mean feature column and σ_x the correspondent standard deviation. These two techniques must always be adapted on the training set and then applied to test set or every new point.

Another important concept is that of **class**. A class is a set of objects which share some properties, for example in this work all the patients who had a cardiovascular event. In supervised approach, the class is represented by a **label** (a common categorical identifier for all the objects belonging to the same class) that can be a positive integer, a character, a string or other types. In the unsupervised approach, conversely, the class is represented by a prototype (an object representative of a specific class). It can either be a real object or just an abstraction and for example it can be computed as the center of gravity of all the objects belonging to a class. Each sample is associated to a class label or a class prototype (Figure 11), therefore if n is the number of dataset samples, the class labels can be represented as a n-dimensional vector. The classification framework can be represented by a simple function:

$$y = f(x)$$

x is the input data (the features vector), f the prediction function and y the output label. During the training, given a set of labeled examples, the prediction function is estimated by minimizing the prediction error on the samples. During the testing, the prediction function is applied to a sample that has been never seen before to give a predicted value as output.

3.2 Dataset Preprocessing Phase

Data quality and quantity of useful information they contain are key factors to determine the learning goodness of a machine learning algorithm. Therefore, data preprocessing is a fundamental phase of a machine learning pipeline. One of the problems to face is the presence of **missing values**. This a normal situation when dealing with biomedical data and patients in real world. Unfortunately, almost every computational tool is not able to manage missing values and hence they must be deleted or imputed. Deletion is a simple choice, but it can be risky, because removing too many samples may lead to unreliable systems. More frequently, missing values are imputed, with various interpolation techniques. For example, missing values can be substituted by their mean value over all the samples or the median or the most frequent one. Another more rigorous method is estimating them with their mean value over the general population. The other fundamental phase of data preprocessing is **dataset partition**. A larger test set gives a more reliable estimate of accuracy, but a larger training set is more representative of how much data is used for the training phase. Further, a single training set is not informative on how much sensitive accuracy is to a specific training sample and therefore how much robust the machine learning algorithm is. Two different solutions to this issue are **holdout model** and **cross validation technique**. Holdout model divides the original dataset into a training dataset and a test dataset, the first one used to train the model and the second one used to estimate its performances. However, another machine learning application step is selecting optimal values for the model's hyperparameters. If the same test set is used more than a time for selecting the model, it will become part of the training data and the algorithm will be more subject to overfitting. Hence, the best way to apply holdout model is by dividing the original dataset into three parts: a **training set**, a **validation set** and a **test set**. The training set is used to adapt different models and the validation set performances are then used to select the model. There is a big advantage in having kept the test set separately and it is that a reliable estimation of the generalization ability of the model can be obtained. Figure 12 shows a schematic representation of holdout model.

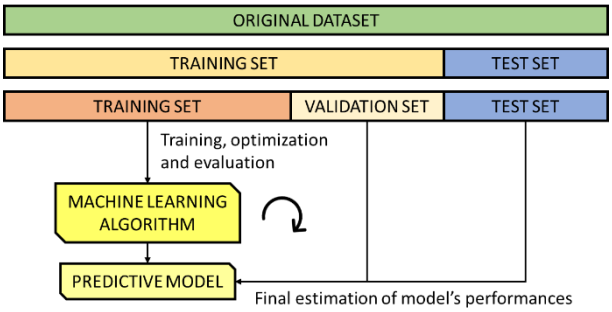

Figure 12 - Holdout cross validation. Validation set is used to repeatedly estimate model's performance after it has been trained using different hyperparameters values. After that, model's generalization error can be estimated on the test set.

The biggest disadvantage of holdout method is that performances estimation is susceptible to the way in which the training set is divided, and it will vary for different data samples. This issue can be addressed by repeatedly partitioning the available data into random partitions: at each iteration, different combinations of training and test sets are randomly selected from the original dataset. This is also called **bootstrap technique** and a schematic representation is shown in Figure 13.

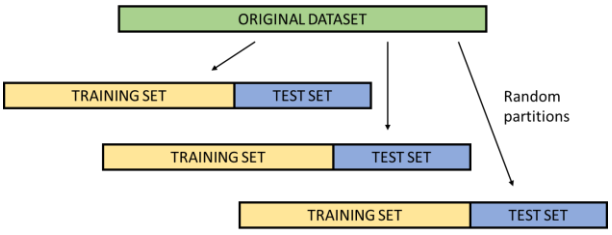

Figure 13 - Bootstrap technique for creating training set / test set partitions.

In k-fold cross validation the dataset is randomly divided into k parts without replacement: k – 1 parts are used to train the model and 1 part is used to test it. The procedure is repeated k times to obtain k models and k related performance estimations. Afterward, the result is a mean over all the k models, therefore performances estimation is less susceptible to data partition then in the holdout model. Since k-fold cross validation is a without replacement resampling technique, it has the advantage that every sample will belong to training dataset only one time and hence it has a lower variance then holdout method. Figure 14 summarizes the concept upon which cross validation is based (here a case with k = 4 is represented). The training dataset is divided into 4 parts and, during 4 iterations, 3 parts are used for training the algorithm and 1 part is used for testing it. Therefore E_i

(algorithm's estimated performances, e.g. classification accuracy) of each part are employed to calculate $\bar{E}$ (mean estimated performance of the model).

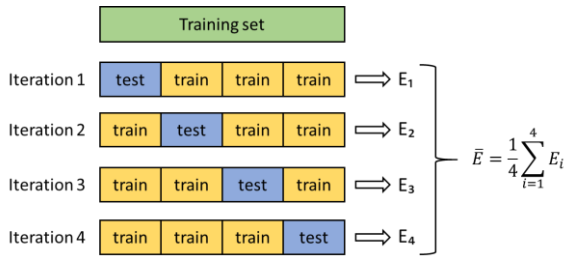

Figure 14 - An example of 4-fold cross validation technique.

K's standard value is 10, and it is reasonable for most applications. Anyway, with small datasets it could be useful to increase K's value, because more training samples will be used in the training phase and therefore a smaller bias on mean performances estimation calculated on each K^{th} part will be generated. However, when K is too high cross validation algorithm has a higher execution time and at the same time a higher variance, because training parts are more like each other. Many alternative versions of cross validation technique exist, like .632 Bootstrap cross validation method (Efron and Tibshirani, 1997).

3.2.1 Features Analysis

This analysis step can be included in the preprocessing phase and it has the aim of improving data representation before training the machine learning system. There are two main approaches to summarize data: **dimensionality reduction (DR)** and **feature selection (FS)**. DR is based on mathematical recombination of the original features and therefore new features are different from the original ones. FS is based on choosing a subset of the original features, hence it takes as input a features vector of m elements and returns a new features vector of n < m elements. FS is equivalent to project the feature space to a subspace of lower dimensions, perpendicular to removed features, which is a subset of the original one (Ferri *et al.*, 1994). DR uses other kinds of projections and the new space is not a subset of the original one, but a recombination of it. Both methods increase the ability of the classifier to well separate different objects belonging to different classes. Table 6 summarizes the main differences between FS and DR.

Table 6 - A comparison between feature selection and dimensionality reduction.

Dimensionality Reduction	Feature Selection
When novel patterns must be classified, all features need to be calculated.	When novel patterns must be classified, only a small number of features need to be calculated (i.e. faster classification).
The measurement units of the original features are lost.	The measurement units of the original features are preserved (this is very important when dealing with biomedical data).
Some methods don't provide good results for all data distributions but work on linear combinations of features.	

Examples of DR are **principal component analysis** and **linear discriminant analysis**. Inside FS techniques, a useful approach to select relevant features from a dataset is by means of **random forest's features importances**. Random forest is an ensemble method for classification and it can measure features' characteristics such as the mean reduction of impurities. This characteristic is computed by all random forest's trees without any assumption on the fact that data are linearly separable or not. About interpretability, it must be mentioned that if two features are highly correlated, one of them may be evaluated as very important, while the other one not. This is a situation that must be considered when we are interested in interpreting features' importances.

3.3 Learning Phase: Models Optimization

It is important to notice that whenever multiple training sets are used, as in cross validation, a learning method is evaluated, not just a learned model. Inside machine learning field, there are two types of parameters: those which are learnt from training data (e.g. the weights of a neural network) and algorithms' parameters or hyperparameters, which are optimized separately (e.g. the number of trees of a random forest). A powerful technique used to optimize hyperparameters is **grid search** and it helps to obtain models' better performances, trying to reach the optimal hyperparameters' combination. Grid search approach is simple: it consists of an exhaustive brute force search in

which a list of values for each hyperparameter is specified. The algorithm, then, evaluates model's performances for each combination, until it reaches the best set. The drawback of this technique is its computational cost. If the hyperparameters' combinations are many, their evaluation may be very expensive. An alternative approach is therefore **random search**.

3.4 Validation and Prediction Phase: Evaluating Classification and Predicting Performance

Assessment of model's performances is needed to set model's parameters, choose the best model between a set of different methods and get an unbiased estimate of the accuracy of a learned model on an unseen test dataset. An **error** happens when a sample is classified as belonging to one class when it belongs to another one. The **error rate** is the percentage of misclassified samples out of the total samples in the test dataset, while the **accuracy rate** is calculated as 100 – error rate. Before presenting the various evaluation metrics, it is convenient to represent the so-called **confusion matrix** (the binary classification version will be considered), a simple square matrix that depicts **true positives** (TP), **true negatives** (TN), **false positives** (FP) and **false negatives** (FN) count (see Figure 15). The definition of positive (P or 1) or negative (N or 0) class is conventional. Usually, the positive class is that in which there is more interest. True positives (or hits) are positive cases correctly identified, true negatives (or correct rejections) are negative cases correctly identified, false positives (or false alarms) are negative cases identified as positive (and represent type I error) and false negatives (or misses) are positive cases identified as negative (and represent type II error). Prediction error and accuracy provide general information about the number of samples that have been incorrectly classified. The error is the sum of all false predictions divided by the number of total predictions:

$$Error = \frac{FP + FN}{FP + FN + TP + TN}$$

Accuracy is the sum of all correct predictions divided by the number of total predictions:

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN} = 1 - Error$$

True positive rate (TPR) and **false positive rate** (FPR) are evaluation metrics particularly useful when dealing with strongly unbalanced classes.

$$TPR = \frac{TP}{P} = \frac{TP}{FN + TP}$$

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN}$$

In tumors' diagnosis, for example, the major interest is in detecting malignant tumors, but at the same time in reducing the number of false positives so that patients are not put into alarm without any reasons. TPR gives information about the fraction of positive samples that have been correctly classified as positives. **Recall** (also called **sensitivity**) is a synonymous of TPR, while **precision** is strictly correlated to other parameters:

$$Precision = \frac{TP}{FP + TP}$$

Specificity (also called **true negative rate**) says how sensitive the classifier is in identifying negative samples:

$$TNR = \frac{TN}{N} = \frac{TN}{FP + TN}$$

The receiver operator characteristic (ROC) curves are useful tools to select classification methods on the base of their performances with respect to FPR and TPR, which are calculated moving the decision threshold of the classifier. In fact, binary classification is often performed by means of a cut-off value or parameter's threshold. For example, if the cut-off value is 0.50, all the samples with output ≥ 0.50 will be classified as positives, while those with output < 0.50 will be classified as negatives. The best cut-off needs to be empirically determined. As Figure 15 shows, the diagonal of a ROC plot can be interpreted as the result of a random guess classifier. Classification models which fall under this diagonal are considered worse than a casual choice.

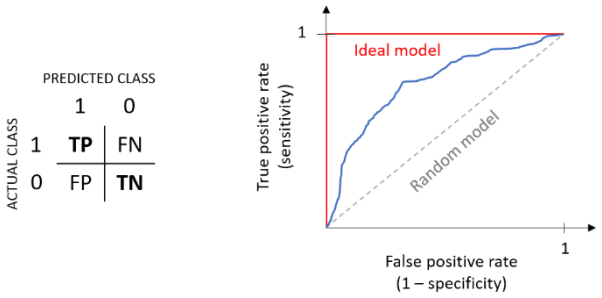

Figure 15 - On the left, a confusion matrix for binary classification. On the right, a ROC curve.

The ideal classifier places itself in the upper left corner of the plot, with TPR = 1 and FPR = 0. On the base of the ROC curve, the area under the curve (AUC) can be calculated. The higher the AUC, the better the classification is. A random classifier has an AUC of 0.50, while an ideal classifier has an AUC of 1.

3.4.1 Underfitting and Overfitting Problems

There are different performance metrics to consider when choosing between different models. The **training error** is the classification error estimated on the training data and it is a measure of how well the model fits the training set. The **test error** is the classification error estimated on the test data, which is different and independent from the training data. It measures how well the model fits new data. **Generalization** capability of a model is described as the goodness with which a learned model can generalize from the data it was trained on a completely new dataset. This concept is linked to **bias** and **variance**, as well as to **underfitting** and **overfitting**. Bias describes how well the average model over all training sets differs from the true model, therefore it can be due to inaccurate assumptions or simplifications made by the model. Variance describes how much models estimated from different training sets differ from each other. These concepts lead to those of underfitting and overfitting. They are two different issues a model can suffer of. The first means that the learned model is too simple to represent all the relevant classes characteristics. It is pointed out by a high training error and a high test error, together with a high bias and a low variance (because the model is consistently bad). The second means that the model is too complex and fits irrelevant characteristics (noise) in the data. It is pointed out by a low training error and a high test error, because the model fits very well training data, but it is not able to generalize. Bias is low (the average model is good), and variance is high (the model changes every time the training set changes). When choosing a classifier, it is important to remember that there is not an inherently better classifier with respect to the others, but assumptions must be made to generalize. Three kinds of errors exist: inherent (unavoidable), bias (due to oversimplifications) and variance (due to inability to perfectly estimate parameters from limited data). To reduce variance a simpler classifier can be chosen, parameters can be regularized or, if it is possible, more training data can be collected. However, a trade-off between bias and variance it is always present.

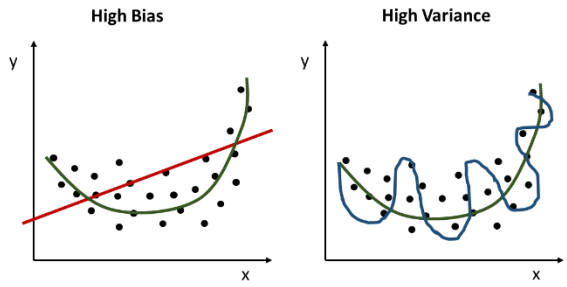

Figure 16 - A schematic representation of the difference between bias and variance. The green line represents the best fitting, while the red line and the blue line represent models with high bias and high variance respectively.

Figure 16 shows the differences between bias and variance. In the first case, the model suffers of underfitting (high bias) and has too few parameters to accurately describe data. In the second case, the model suffers of overfitting (high variance), has too many parameters and depends too much from the training dataset.

3.5 Three Classifiers Examples

3.5.1 K-Nearest Neighbor Classifier

K-nearest neighbor (KNN) classifier is a simple machine learning algorithm and it is a typical example of lazy learning system, because it does not learn a decision function from training data. Moreover, it can be classified as a non-parametric model: it is not possible to characterize it by a fixed number of parameters, since this number increases based on the number of training data. More precisely, KNN belongs to a subgroup of non-parametric models called instances-based learning models. The main steps of KNN are:

1. Choose K and a distance metric (ex. Euclidean distance)
2. Find K elements nearest to the subject you want to classify
3. Assign the label to the subject on the base of a majority voting among the neighbors

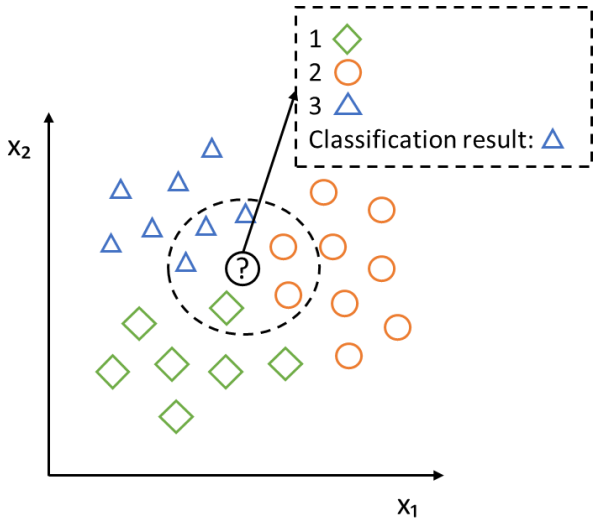

Figure 17 - A practical example of classification by means of k-nearest neighbor classifier.

Figure 17 illustrates how a new datapoint (represented by "?" symbol) is assigned to triangle class on the base of a majority voting among its k nearest neighbors, which are six in this specific case. The advantage of this memorization-based approach is that the classifier immediately adapts itself while collecting new data. The disadvantage is that computational complexity required to classify new instances grows linearly together with the increasement of training samples. Therefore, memorization space can become a problem if the dataset has huge dimensions. Different KNN implementations exist, and some of them use efficient data structures like KD-trees (Freidman, Bentley and Finkel, 1977). Choosing K might be tricky and the right K balances underfitting and overfitting of the classifier. Another important parameter to choose is the distance metrics, that must be appropriate for the current dataset. The simplest measure is the Euclidean distance:

$$d\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right) = \sqrt{(x^{(i)} - x^{(j)})^2}$$

where $x^{(i)}$ and $x^{(j)}$ are two datapoints belonging to the training dataset. However, when using the Euclidean distance, it is important to standardize data so that each feature can equally contribute to distance. Another distance that can be used is Minkowski, a generalization of Euclidean and Manhattan distances:

$$d\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right) = \sqrt[p]{\sum_k |x_k^{(i)} - x_k^{(j)}|^p}$$

KNN is very susceptible to overfitting problem: this happens when features space becomes too sparse and nearest neighbors are too far to give a stable estimation. A possible way to solve it is by reducing features space.

3.5.2 Support Vector Machine Classifier

Support vector machine (SVM) classifier can be considered an extension of perceptron, which belongs to the first machine learning algorithms historically described for classification (Rosenblatt, 1958; McCulloch and Pitts, 1990). However, it minimizes misclassification errors, while the optimization objective of SVM is maximizing the margin. The margin is defined as the distance between the separation hyperplane (decision function) and the datapoints which are nearest to this hyperplane (the so-called support vectors). Figure 18 shows graphically this concept.

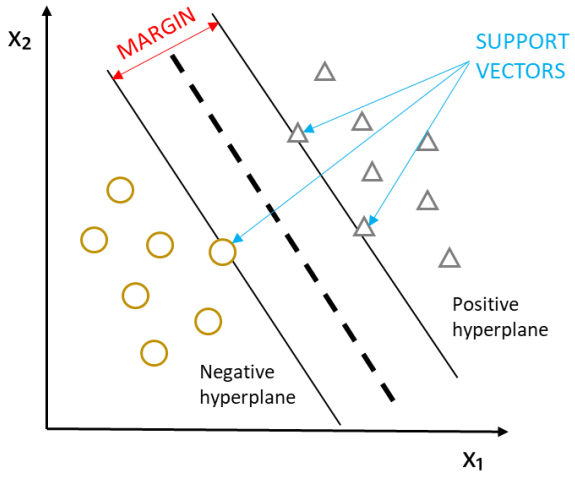

Figure 18 - A graphical explanation of the concept upon which support vector machine classification is based: the margin maximization.

Decision functions with large margins possess a lower generalization error, while models with small margins are more susceptible to overfitting. The positive and negative hyperplanes, which are parallel to decision function can be described as follows:

$$w_0 + \boldsymbol{w}^T \boldsymbol{x_{pos}} = 1$$

$$w_0 + \boldsymbol{w}^T \boldsymbol{x_{neg}} = -1$$

If these functions are subtracted, we obtain:

$$\boldsymbol{w}^T(\boldsymbol{x_{pos}} - \boldsymbol{x_{neg}}) = 2$$

Normalizing the equation for **w** length, which is defined as follows:

$$||\boldsymbol{w}|| = \sqrt{\sum_{j=1}^{m} {w_j}^2}$$

we obtain:

$$\frac{\boldsymbol{w}^T(\boldsymbol{x_{pos}} - \boldsymbol{x_{neg}})}{||\boldsymbol{w}||} = \frac{2}{||\boldsymbol{w}||}$$

The left side of the equation can be interpreted as the margin (the distance between the positive and the negative hyperplanes) that must be maximized. Therefore, the objective function of SVM becomes the maximization of $\frac{2}{||\boldsymbol{w}||}$ under the constrain that samples are correctly classified, that can be expressed as follows:

$$w_0 + \boldsymbol{w}^T\boldsymbol{x}^{(i)} \geq 1 \; if \; y^{(i)} = 1$$

$$w_0 + \boldsymbol{w}^T\boldsymbol{x}^{(i)} < -1 \; if \; y^{(i)} = -1$$

These two equations state that all negative samples must stand on one of the sides of the negative hyperplane and all the positive ones on the other side of the positive hyperplane. Practically, minimizing $\frac{1}{2}||\boldsymbol{w}||^2$ is simpler, by means of quadratic programming. A linear version of SVM exists (**soft margin classification**), which employs the so-called slack variable ξ. The reason of introducing ξ is that linear constrains must be reduces in the case of not linearly separable data, to consent optimization convergence. Slack variable for positive values is simply added to linear constrains:

$$\boldsymbol{w}^T\boldsymbol{x}^{(i)} \geq 1 \; if \; y^{(i)} = 1 - \xi^{(i)}$$

$$\boldsymbol{w}^T\boldsymbol{x}^{(i)} < -1 \; if \; y^{(i)} = 1 + \xi^{(i)}$$

Hence, the new objective function becomes:

$$\frac{1}{2}||\boldsymbol{w}||^2 + C(\sum_i \xi^{(i)})$$

Tuning C, penalization for wrong classification can be controlled. Big C values correspond to big penalizations of errors, while small C valued correspond to small penalizations. Therefore, C can

be used to control the margin width and to optimize bias-variance trade-off. If the dataset to be classified is just too hard, a nonlinear version of SVM (**kernel SVM**) exists to solve it. The general idea is that the original input space can always be mapped to some higher-dimensional feature space where the training set is separable, creating new nonlinear combinations of the original features. Figure 19 intuitively shows this concept.

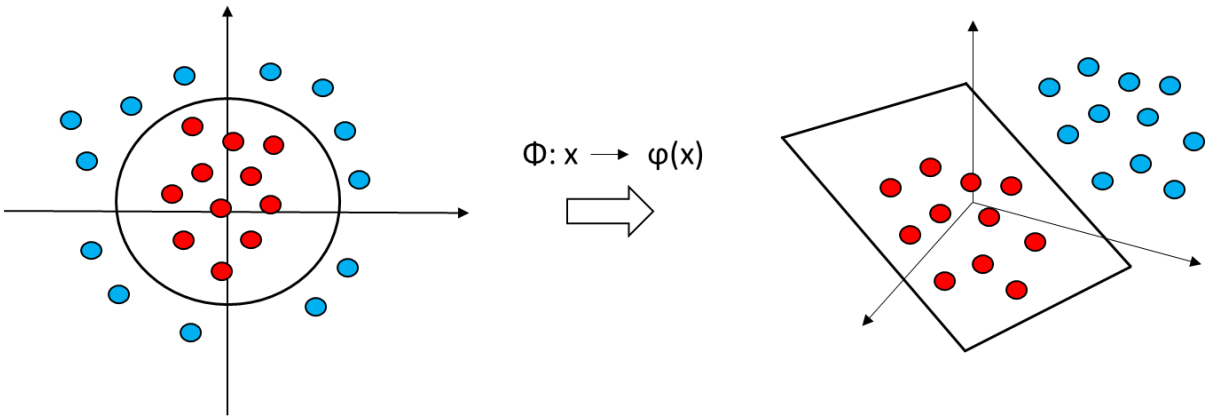

Figure 19 - A graphical representation of the idea upon which kernel support vector machine are based. Φ is the kernel function. On the left, the original features space is represented. On the right, the new features space is shown, after kernel transformation.

In practice, the training dataset is implicitly mapped into a higher-dimensional space using a kernel function Φ (with a minimal effect on computation time, made possible by these kernel functions). Then, a linear SVM is trained on this transformed space. When new unseen data must be classified, they are transformed by means of Φ function and classified by linear SVM previously trained. Kernel must be chosen by the user (ex. polynomial kernel, radial basis function kernel, sigmoid kernel…). Radial basis function (RBF) kernel or gaussian kernel is one of the most popular.

$$k\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right) = e^{-\frac{||\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}||^2}{2\sigma^2}}$$

$\gamma = \frac{1}{2\sigma^2}$ is a parameter that needs to be optimized by the user, together with C parameter (trade-off between misclassification of training samples and simplicity of the decision surface). Kernels can be interpreted as similarity functions between couples of samples. Similarity score is always between 0 (very different samples) and 1 (very similar samples). Therefore γ defines the influence of a single training sample and plays an important role in overfitting. The advantages of using SVM are that the kernel framework is very powerful and that these classifiers work very well in practice,

even with a small training set, because most of the reliability depends on support vectors. The disadvantages are the computational cost of this method in the case of large-scale problems and the fact that tuning kernel's parameters might be tricky.

3.5.3 Decision Tree and Random Forest Classifiers

Decision tree (DT) classifiers are powerful models if we are interested in interoperability. DT classifies data taking decisions on the base of answers to a series of questions. A simple example is represented in Figure 20.

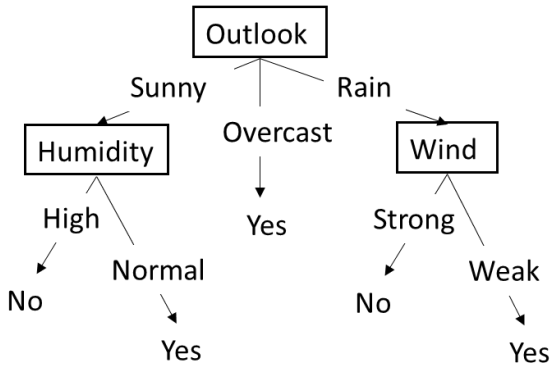

Figure 20 - A simple decision tree to decide if someone should go out or not.

DTs have a flowchart-like structure containing nodes, branches and leaves. Each node specifies a test involving an attribute (i.e. one of the features) and every branch descending from a node represents one of the possible outcomes of the test (i.e. one of the possible values of the corresponding feature). Leaves are nodes with the final class label. Classifying an instance means performing a sequence of tests, starting from the root node and terminating with a leaf one. Automatic ways to represent data in the form of a DT exist and are based on induction. Top Down Induction of Decision Trees is a possible family of methods to induct DTs from a dataset. One of the algorithms of the family is ID3 (developed by Quinlan), an iterative technique inside a top-down approach. The pseudo-code to construct a DT (T) from a learning set (S) is:

- If all examples in S belong to the same class C, then make a leaf labeled C
- Otherwise
    - Select the "most informative" attribute (A)
    - Partition S according to A's values

- Recursively construct subtrees T1, T2, … for the subsets of S (one for each value of A)

Selecting the most informative attribute means selecting the one that partitions the dataset into subsets as homogeneous as possible in terms of class labels. To classify an object, a certain information (I) is needed and after applying attribute A, only a residual information (I_{res}) is required to classify that object. The gain is defined as the difference between the initial information and the residual information:

$$Gain(A) = I - I_{res}(A)$$

The most informative attribute is the one that maximizes the gain. Entropy is defined as the averaged amount of information needed to classify an object:

$$I = -\sum_{p} p(c) log_2 p(c)$$

where p(c) is the proportion of samples belonging to class c. If all the samples belong to the same category, the entropy is minimum (0), while if the samples are equally mixed (1/c samples for each class), the entropy is maximum (1). After applying attribute A, S is partitioned into subsets according to A's values v. Residual information is equal to the weighted sum of the amounts of information for the subsets:

$$I_{res} = -\sum_{v} p(v) \sum_{c} p(c|v) log_2 p(c|v)$$

A possible measure of attributes' homogeneity is **information gain**. The attribute with the highest gain is chose as the most informative. A limitation of this approach is that residual information favors attributes with many values. One possible solution to this issue is using a corrected measure: **information gain ratio**. It is obtained divided the gain of A by its information:

$$GainRatio(A) = \frac{Gain(A)}{I(A)} = \frac{I - I_{res}(A)}{I(A)}$$

Another sensible measure of impurity is **Gini Index**:

$$Gini = \sum_{i \neq j} p(i) p(j),$$

where p(i) is the proportion of examples of a class and i and j are the classes. Gini is the measure of the initial information. After applying attribute A, the resulting Gini Index is:

$$GiniGain(A) = \sum_v p(v) \sum_{i \neq j} p(i|v)p(j|v),$$

where p(i|v) is the information of a subset provided that a specific path has been followed. GiniGain(A) is the measure of the residual information. To summarize, Gain(A) is based on entropy, GainRatio(A) accounts for the heterogeneity between features and GiniGain(A) depends on the probability of classifying an object wrongly. Strengths of DTs are their easiness to be implemented and understood, while their weaknesses are the tendency to overtraining and the necessity to data discretization, because they produce complex decision regions.

Random forest (RF) classifier is an ensemble method specifically designed for DT classifiers. The idea that drives to use an ensemble of learning methods is that by combining more weak learners it is possible to build a stronger learner with a better generalization error and less susceptible to overfitting. The main steps of RF algorithm are:

1. Randomly choose n training set samples with replacement (initial bootstrap sample)
2. Grow a decision tree from the previous extracted sample. For each node:
    a) Select randomly d characteristics without replacement
    b) Divide the node based on the characteristic which gives the best division (through the objective function)
3. Repeat k times steps 1. and 2.
4. Put together each decision tree prediction to assign the final label on the base of a majority voting

In the second step, instead of evaluating functions on all features to determine the best split in each node, only a random subset is considered. RFs do not offer the same interoperability level of DTs, but they present great advantages: hyperparameters do not require a strong optimization and the model is resistant with respect to noise. The only hyperparameter to set is the number of decision trees to use.

4. Machine Learning and Biomedical Engineering

Machine learning (ML) can be applied in almost every computing task and it has shown outstanding performances in the fields of speech recognition, natural language processing, computer vision and robot control. Concerning biomedical applications, the recent and rapid developments in advanced computing and imaging systems have opened the way to a new research dimension together with the increasing available data, which requires specific ML approaches to successfully exploit its content (Park, Took and Seong, 2018). The huge field of artificial intelligence (AI) may optimize the care journey of chronic disease patients, find personalized precision therapies for complex illnesses, reduce medical errors due to human cognitive biases, speeding up diagnostic processes, and make subject enrollment in clinical trials more effective. As a simple example, natural language processing can be used to analyze the expanding scientific literature and summarize different electronic medical records together. Figure 21 shows various AI applications in everyday life.

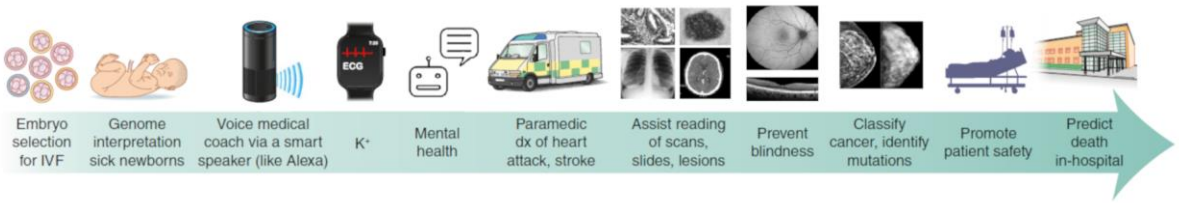

Figure 21 - Examples of artificial intelligence applications in everyday human life (dx: diagnosis, IVF: in vitro fertilization, K^+: potassium blood level).

Successful AI applications already exist in image analysis for radiology, pathology and dermatology (Miller and Brown, 2018). Diagnostic confidence of these tools never reaches 100%, however combining machines results and physicians experience can enhance overall performances.

4.1 Historical Artificial Neural Networks Applications

The most famous ML method is probably that of *artificial neural networks* (ANNs). They are flexible mathematical models that can understand complex relationships between variables contained in big datasets, gradually modifying a series of coefficients based on the errors encountered during the model building process. Simple ANNs have been employed in medicine

since the 1990s to **read electrocardiograms** (Willems *et al.*, 1991), **diagnose myocardial infarction** (Baxt, 1991) and as a **predictive mean for the intensive care unit length of stay following cardiac surgery** (Tu and Guerriere, 1993).

In the first case, a large international study was performed to compare the performances of nine computer programs developed for electrocardiograms (ECGs) analysis with those of eight cardiologists in reading and interpreting ECGs in 1220 clinically validated cases of different cardiac pathologies. What results showed was that the percentage of correctly classified ECGs by computer programs (median, 91.3%) was lower than the cardiologists one (median, 96%; $P<0.01$). However, the performance of the best programs was almost pair to that of the most accurate cardiologists. This study highlights the promising results which ML can obtained and the fact that it can be a strong help for the physician, whose review of computerized reports is always essential. Computerized electrocardiography has a big margin for improvement, for example combining different programs' results to increase diagnostic accuracy.

The second paper was intended to validate prospectively the use of an ANN to identify the presence of myocardial infarction in 331 patients arriving to an emergency department with anterior chest pain and comparing its results with that of physicians. The physicians had a diagnostic sensitivity of 77.7% (95% confidence interval, CI, 77.0% to 82.9%) and a diagnostic specificity of 84.7% (95% CI, 84.0% to 86.4%), while the ANN a sensitivity of 97.2% (95% CI, 97.2% to 97.5%; $P=0.033$) and a specificity of 96.2% (95% CI, 96.2% to 96.4%; $P<0.001$). Figure 22 shows the network used. Each ANN is made of at least three layers: the input layer, the hidden layer and the output layer. Every layer has a certain number of processing units that are connected between them by weights (i.e. coefficients). The number of hidden units together with the number of layers of hidden units were chosen by trial and error.

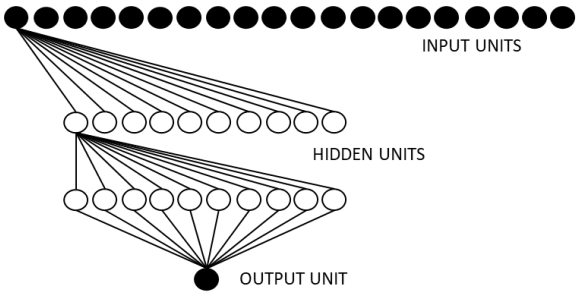

Figure 22 - Scheme of the 20x10x10x1 backpropagation network used in the paper. Circles represent the processing units, while lines represent the connection weights, the place in which the network stores the learnt information.

The inputs were selected from the presenting symptoms, the past clinical history, the physical and laboratory tests of enrolled patients. Before this network can be considered a legitimate aid to physicians during clinical diagnosis, its performance must be prospectively validated using a larger dataset. Nevertheless, this is a promising result. It indirectly shows that a kind of network like that may identify relationships in clinical data which have not been elucidated yet and cannot be appreciated by clinicians. Another advantage is that it applies non-linear statistics, so it may potentially solve more complex problems.

The third paper had the objective of building a computer system to predict patients' intensive care unit length of stay after cardiac surgery. The need was to improve the use of existing intensive care units and resources by means of better scheduling and optimization of patients and staff. An ANN was trained on 713 patients using 15 parameters. A length of stay greater than 2 days was considered prolonged. The model was then independently validated on 696 patients, being able to stratify them into three risk groups (low, intermediate and high prolonged stay) according to different frequencies of stay. This was possible because the output of the network was analyzed as a probability and discretized by two different thresholds. The ANN was also evaluated by calculating the area under the receiver operating characteristic (ROC) curve (AUC) on both the training set (0.7094 with a standard error, SE, of 0.0224) and the test set (0.6960 with a SE of 0.0227). The conclusion of this study is that an ANN can successfully be employed as predictor of patients' length of stay in an intensive care unit, but if a similar kind of system can be used in a hospital by real doctors remains to be determined. The reason is the black box nature of ANN and the impossibility of explicitly determining the relationships between clinical variables, necessary to understand the prediction. Another interesting suggestion present in this article is the comparison

between ANNs and other traditional statistical techniques, such as logistic regression, as Table 7 shows. The authors believe that ANNs will not replace other traditional methods unless they are easier to implement and perform more accurately.

Table 7 - Comparison between ANNs and logistic regression as a predictive instrument in the clinical field.

Artificial neural network	Logistic regression
It better identifies complex, non-linear relationships between variables.	It better identifies simple, linear relationships between variables.
Developer does not require big knowledge about ANNs.	Developer needs to have a substantial background in statistics.
Black box nature: relationship between input variables and predicted output cannot be explained.	Relationships between input variables and predicted output clearly identified.
Computationally expensive.	Computationally simple to implement.
It can handle fuzzy and missing data.	It assumes data is complete and accurate.
It requires the trained model to be used.	It can be easily used also with a hand calculator.
It is not known if clinicians accept this technique.	It is familiar to clinicians and accepted by them.

4.2 The Most Recent Deep Learning Applications

The most novel field of machine learning today is the so-called *deep learning* (DL). It was first conceived for image classification tasks, trying to replicate the mammalians' visual cortex. It has incredibly improved the state of the art in speech recognition, visual object recognition, object detection and many other fields such as drug discovery and genomics (LeCun, Bengio and Hinton, 2015). It has been applied since 1990s, but only recently supercomputer speed gave the possibility of exploring massive dataset with it. The typical structure of a DL algorithms is a cascade of many hidden layers made of locally connected units, for feature extraction and transformation. A comparison between traditional machine learning workflow and deep learning one is represented

in Figure 23. A limit of traditional machine learning is the need of data processing in its raw form, because it requires a big engineering step to design a feature extractor so that the classifier has in input a suitable feature vector which can recognize. In addition to this, hand-crafted features are highly application-dependent. On the contrary, DL is a representation-learning method composed of different levels of non-linear units which perform increasing abstract transformations on the raw data. These levels form a hierarchy of concepts. The key aspect of DL is that each layer is not designed by human engineers but is learned from data using a general-purpose learning procedure.

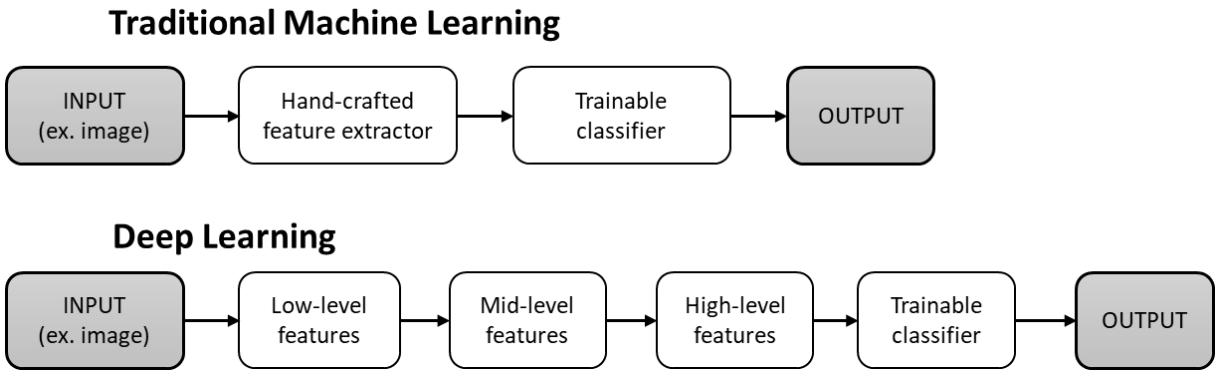

Figure 23 - Traditional machine learning vs deep learning workflow. Deep learning tries to learn rich hierarchical representations automatically by means of multiple feature learning stages.

DL and traditional ML have been employed in cognitive diagnostic, chronic disease management, and electronic medical records applications.

Convolutional neural networks (CNNs) were applied in **keratinocyte carcinoma and melanoma detection** and gave better results than dermatologists (Esteva *et al.*, 2017). CNNs is a feed-forward artificial neural network whose connectivity pattern between neurons is inspired from the human visual cortex organization. They are made of different layers with different levels of abstraction with respect to the classification task. Fine-grained variability in the appearance of skin lesions makes them difficult to be automatically classified. CNNs showed promising results dealing with highly variable tasks across many fine-grained object categories. Esteva et al. used a dataset of 129450 clinical images (a lot bigger than previously used ones in this field) to train a CNN that outperformed 21 dermatologists with an AUC over 91%. This method could be easily included in mobile devices and therefore potentially provide low-cost universal access to vital diagnostic care. In fact, the system, since it was trained on a very big dataset, has an incredible generalization property. It can be used with photographic images (for example smartphone images), because it

was pre-trained on 1.28 million general images to make classification robust to photographic variability. This example shows the promising value of DL, that is agnostic to the type of image data and can be adapted to other medical fields, such as ophthalmology, otolaryngology, radiology and pathology.

DL was applied to **predict healthcare trajectories from electronic medical records (EMRs) data** (Pham *et al.*, 2017). EMRs are tools intended for documenting and sharing patients' history about hospitalizations, diagnoses, interventions, laboratory tests and clinical reports. They represent an incredible data source for computational models. In this work, an end-to-end deep dynamic neural network called DeepCare is presented, which reads medical narratives, stores previous pathology history, infers current pathology state and predicts future medical outcomes. DeepCare faces and solves four big challenges of EMRs: the dependency of future illness and care from patient's history, the difficulty of representing admission information, the intrinsic episodic nature of medical records together with their irregular time of length and the ambiguous interactions between disease progression and interventions. It is built on long short-term memory, a recurrent neural network with memory cells to store experiences. Memory is also controlled by a forgetting unit. The model was validated on two different cohorts (diabetes and mental health).

4.3 Machine Learning Applications in Medical Diagnosis and Treatment

The two fundamental processing tasks of healthcare are diagnosis (classification of cases through history and current examination) and treatment (planning and delivery of therapy with a desired outcome). These processes involve hypothesis generation, hypothesis testing and action. Machine learning techniques can help hypothesis generation and testing (Panch, Szolovits and Atun, 2018). They are not based on a priori assumptions about the distribution of the data, can reveal previously hidden relationships between variables and find new patterns and can incorporate many more variables than traditional statistical techniques leading to more generalized models. Some research examples exist in diagnosis and prediction of future events.

In the field of waveforms analysis, an attempt to apply machine learning in the diagnosis phase has been made to **detect surface and age-related differences in walking from a single wearable inertial sensor** (Hu *et al.*, 2018). In this study data from an accelerometer, gyroscope and magnetometer collected by an inertial motion unit were used to feed a DL network with long short-

term memory unit (fully supervised), which can learn the temporal dynamics of sequential data and are hence suited for learning time series data. Data were binned into smaller segments. The dataset comprised 17 older and 18 young healthy adults, who were made walking over flat and uneven brick surfaces wearing the inertial motion unit. 90% of the data was used to train the network and 10% to test it. Four different models using different inputs were compared: the fully trained model, which used all 9 channels from every sensor in the inertial motion unit, accelerometer signals alone, gyroscope signals alone and magnetometer signals alone. The fully trained model outperformed all the others, with an AUC value of 0.97 in the case of surface detection and 0.96 in the case of age detection ($p \leq 0.045$), an accuracy of 96.3% and 94.7%, a precision of 96.4% and 95.2%, a recall of 96.3% and 94.7% and a f1-score of 96.3 and 94.6% respectively. These results show that this method, with further learning, may be used to facilitate identifying and intervening on fall risk. The current method is innovative in two aspects: this kind of network does not need to be fed with complex data processing and feature extraction, but only with the inertial motion unit outputs from which, then, it learns useful features to make predictions autonomously. Therefore, DL methods provide an automated and scalable approach, that can potentially be applied to population-based surveillance of everyday walking behavior. The second innovation is that the difference between the tasks was subtle to detect.

For what it concerns prognosis and prediction, machine learning has already been applied to **cardiovascular event prediction** (Ambale-Venkatesh *et al.*, 2017). In this study, random survival forests were used to predict six cardiovascular outcomes in comparison to standard cardiovascular risk scores. 6814 participants aged 45 to 84 years, from 4 ethnicities and 6 centers across USA were included. 735 variables from imaging and non-invasive tests, questionnaires and biomarker panels were collected. Random forest was used to detect the top 20 predictors of each outcome. Imaging, electrocardiography and serum biomarkers were situated among the best predictors, differently from traditional risk factors. Age remained the most important predictor for all-cause mortality. This study highlights that machine learning methods such as random forest may lead to great insights regarding subclinical illness markers without a priori assumptions of interconnection. Moreover, machine learning methods appear well-suited for meaningful risk prediction in large-scale epidemiological studies, in fact random forest provided better predictions with respect to standard risk scores. These techniques could help doctors in the specific use of variables for specific event prediction and in thinking about new strategies to prevent cardiovascular disease outcomes.

Potentially, these methods could be applied retrospectively as a means of diseases mechanism investigation and hypothesis generation.

In the field of image analysis through artificial intelligent systems, many applications exist. An example is the deployment of DL for **diagnosis and referral of retinal disease** (De Fauw *et al.*, 2018). In this work, a DL architecture was applied to a clinically heterogeneous set of three-dimensional optical coherence tomography scans from patients with eye diseases. The performance of the system exceeded that of experts in making a referral recommendation in a wide range of sight-threatening retinal diseases (5.5% error rate against 6.7% and 6.8% of the two best retina specialists). The training was performed on 14884 scans. Automated diagnosis of a medical image faces two main challenges: technical variations in the image acquisition process (due to the use of different devices, presence of noise etc.) and inter patients' variability with respect to pathology manifestation. Existing DL approaches use a single end-to-end black box network with the disadvantage of requiring millions of images to be trained. In contrast, in this study, the developed framework decoupled the two issues and solved them independently. Another improvement was the detection of ambiguous regions in optical coherence tomography image segmentation. This problem was solved by training multiple instances of the segmentation network and consequent hypotheses in a similar way to multiple human experts. Multiple hypotheses were then displayed as a video with the ambiguous regions clearly visible.

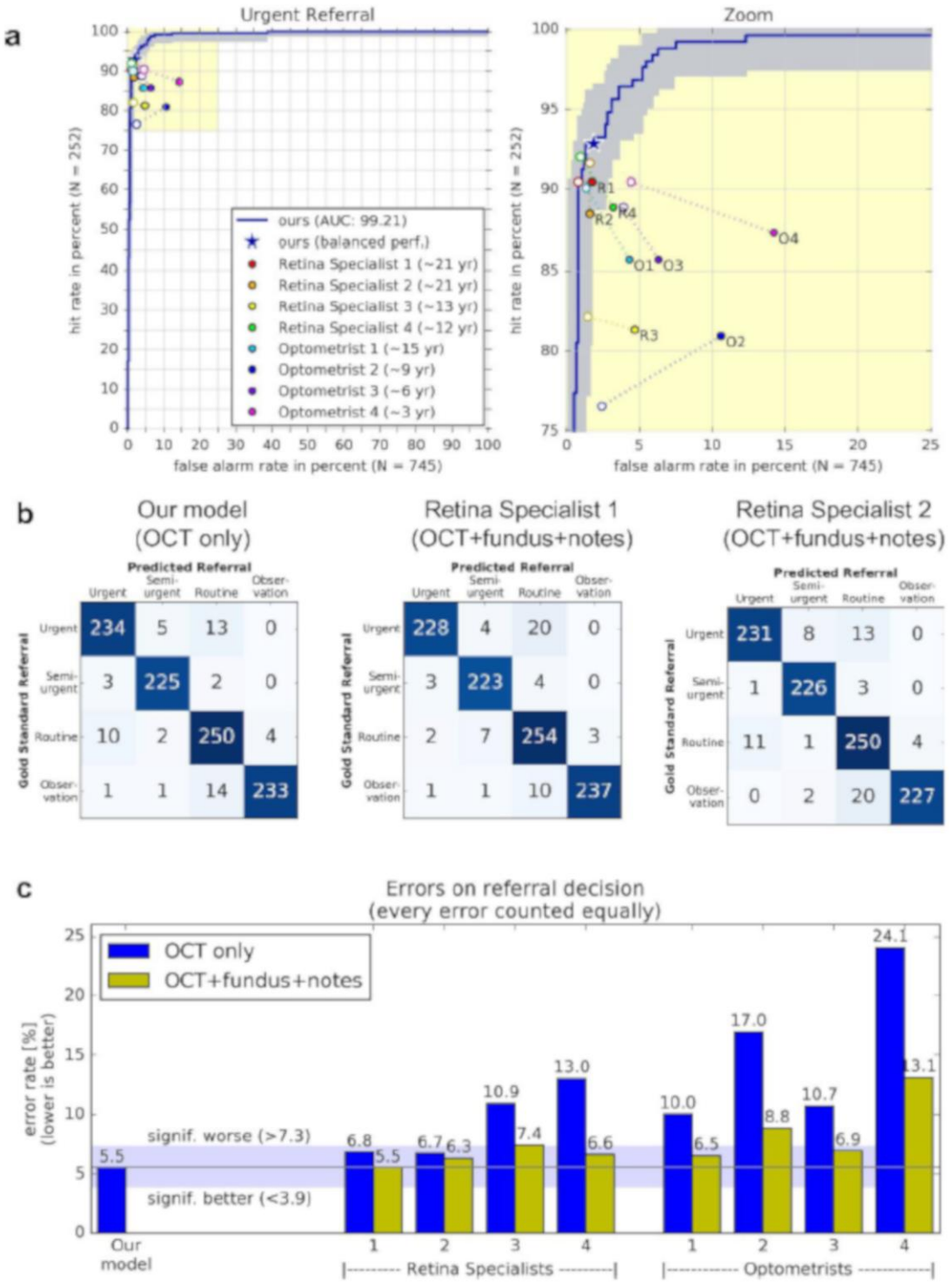

*Figure 24 - Results on the patients' referral decision. (**a**) ROC curves for urgent referrals with respect to other referrals. (**b**) Confusion matrices with patients' number for referral decision in the case of the system developed in this work and the results of the two best specialists. (**c**) Total error rate.*

Figure 24 summarizes the principle results of the paper. The performances are represented on an independent test set of 997 patients (252 urgent, 230 semi-urgent, 266 routine, 249 observation only). The benefits of this work are multiple. The black box issue of artificial intelligence, as an impediment in healthcare application, was limited. The created framework closely matched the clinical decision-making process, allowing clinicians to inspect and visualize an interpretable segmentation, rather than simply read a referral suggestion. This system can potentially be used in clinical training, where medical professionals must learn to read medical images.

About prognosis and prediction, an interesting study is that of **federal learning of predictive models from federated electronic health record** (Brisimi *et al.*, 2018). In an era of "big data" computationally efficient and privacy-aware tools are required, especially in healthcare field, where huge amounts of data are stored in different places and owned by different institutions. Therefore, centralized algorithms are not practicable. They must be substituted by decentralized computational scalable methodologies. In this paper the soft margin regularized support vector machine classifier was employed to solve a binary supervised classification problem to predict hospitalizations for cardiac events. A distributed algorithm was developed with an iterative cluster primal dual splitting to face the large-scale problem in a decentralized fashion. This is a situation in which data reside with many agents (institutions or patients), no raw data get exchanged and the agents collaborate to jointly learn the classifier. The new learning decentralized method is hence flexible and scalable. The entirety of electronic health record was used and existing risk metrics such as Framingham risk score were thus improved.

Support vector machine (SVM) was applied to magnetic resonance mapping of white matter neuronal water content for **prediction of major depressive disorder**, facing with success the issue of phenotypic dimensionality and depression activity markers paucity (Schnyer *et al.*, 2017). SVM obtained a classification accuracy of 74%, showing that in vivo diffusion tensor magnetic resonance imaging can accurately diagnose major depressive disorder. Another interesting result is that prediction is strongest when only right hemisphere white matter is considered. 52 patients with DSM-IV major depressive disorder and 45 healthy control participants were included in the study.

Unsupervised learning was applied to the analysis of patients with **heart failure with preserved ejection fraction** (Shah *et al.*, 2015). This is an attempt toward precision medicine, because this

illness is a phenotypically heterogenous condition with the involvement of different weak genetic factors, without a proven therapy. Therefore, the application of unsupervised learning had the objective of trying to find the different pathophysiological processes characterizing the patients. Unbiased clustering analysis using dense phenotypic data (the so-called phenomapping) was employed to identify different patients' categories. A regularized form of model-based clustering was employed, with multivariate Gaussian distributions to separate each patient's cluster based on the means and standard deviations of each feature. 397 subjects were included in the study, considering detailed clinical, laboratory, electrocardiographic and echocardiographic phenotyping. Starting from 67 features, after removal of correlated ones, 46 features remained and were used. Although all patients had the same diagnosis, they were divided into 3 distinct groups after the clustering. These 3 groups differed in clinical characteristics, cardiac structure and function, invasive hemodynamics and outcomes. Hence, this kind of technique resulted in novel classification of heart failure with preserved ejection fraction, representing a new method to classify heterogeneous clinical syndromes with the ultimate objective of defining homogeneous patients' subgroups which can be treated in the same way. In addition to this, DL algorithms have shown that pre-training them with an unsupervised approach can markedly improve the performance of subsequent supervised classification steps.

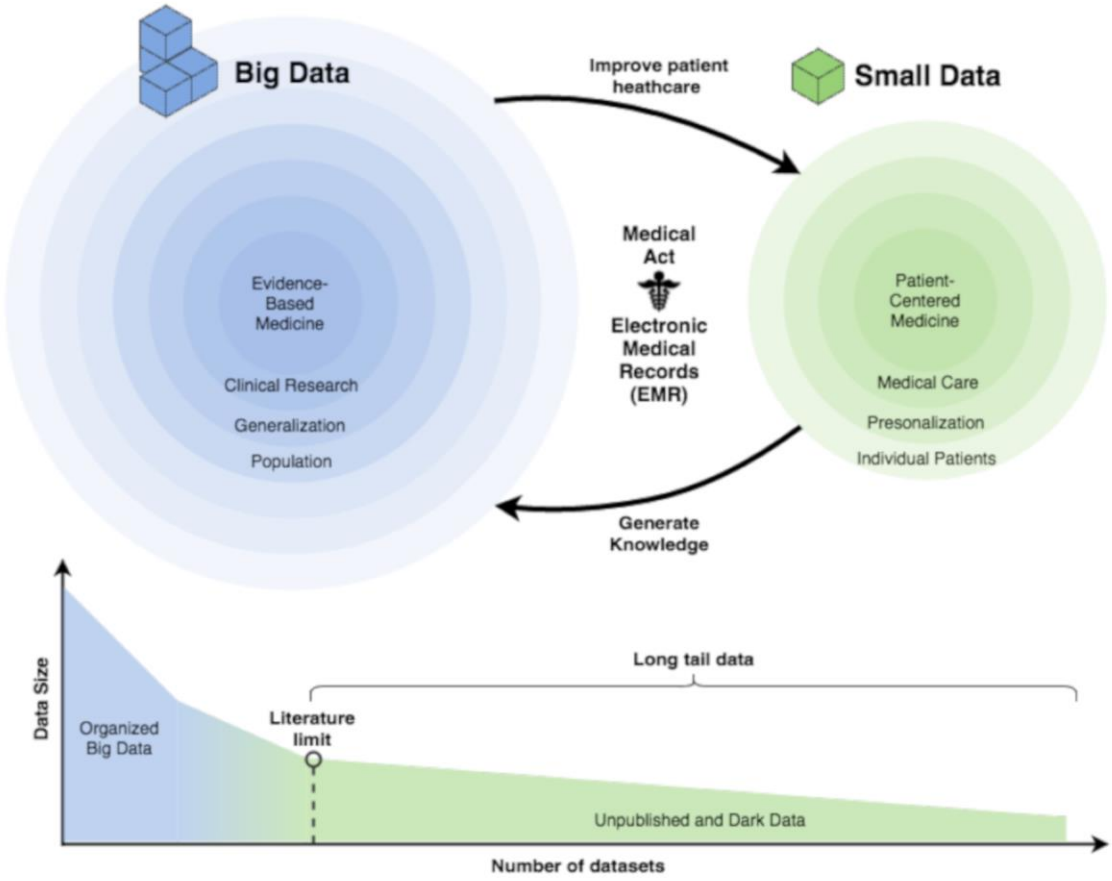

Figure 25 - A possible graphical representation of the relationship between big data and small data in the learning healthcare system.

Figure 25 summarizes the relationship between big data and small data in healthcare and the need to integrate them in a single vision that can accounts both for evidence-based medicine and patient-centered medicine. Moreover, there are a lot of unpublished datasets still not used which can be a potential treasure for big data.

4.4 Machine Learning Applications in the Field of Basic Sciences

ML has also been applied to the field of **basic sciences**, such as computational biology. Thanks to recent advances in high-throughput sequencing technologies, large biological datasets have been made available to the scientific community. Moreover, internet web services expanded and enabled scientists to put a lot of data online. As a result, novel ways to interrogate, analyze and process data were born, to infer knowledge about molecular biology, physiology and biomedicine.

Computational biology and bioinformatics are characterized by huge quantity of data that cannot be handled manually. Besides, expert knowledge in these field is incomplete and inaccurate and there are many exceptions to the general cases. The ability of ML to automatically identify patterns in data is particularly important in these cases.

DL has been applied for **regulatory genomics** (Angermueller *et al.*, 2016). Conventional approaches relate sequence variation to changes in molecular traits. One method is training models that use variation between regions within a genome, as in Figure 26, A. Splitting the sequence into windows that are centered on the trait of interest, tens of thousands of training examples are generated and prediction is challenging. DL presents many advantages: it can operate on the sequence directly, without requiring predefined features and it can capture nonlinear dependencies and interaction effects in the sequence. A convolutional neural network architecture, as represented in Figure 26, B, allows to reduce the number of parameters that the model has as inputs. The key gain of this approach is the ability to train the model on larger sequences. The first convolutional layer searches for motifs along the input sequence, while the second convolutional layer reduces the input dimensions. Any additional layers can better model interactions between the previous identified motifs.

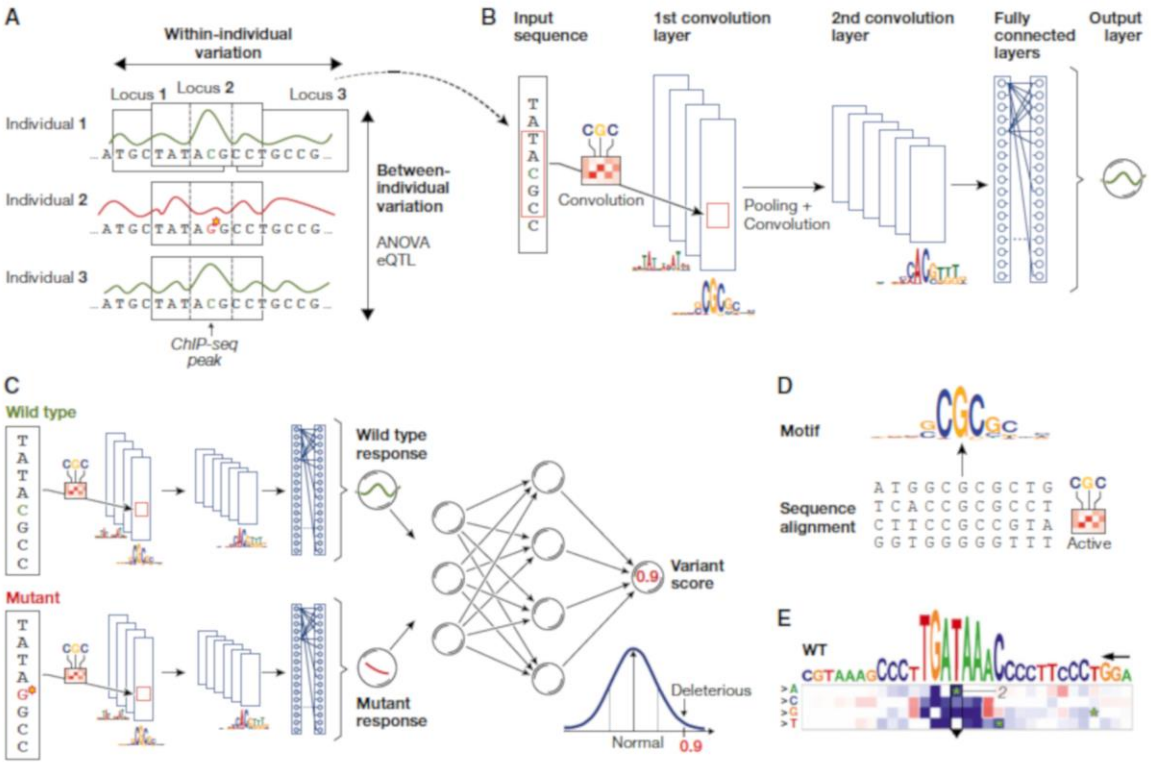

*Figure 26 - Principles of using neural networks for predicting molecular traits from DNA sequence. (**A**) DNA sequence and the molecular response variable along the genome for three samples. (**B**) 1-dimensional convolutional neural network for the prediction of a molecular trait given a raw DNA window. (**C**) Neural network predicts wild-type and mutant sequence by means of a response variable used as input to an additional network that discriminates normal from deleterious variants. (**D**) A convolutional filter that acts aligning genetic sequences. (**E**) Mutation map of a sequence window.*

The motifs identified by the model can be visualized as heatmaps as in Figure 26, D.

Another ML application in computational biology is represented by the **prediction of bitterness and sweetness of chemical compounds** by means of a random forest classifier (Banerjee and Preissner, 2018). In this work a ML model based on molecular fingerprints was developed and validated to discriminate between bitter and sweet molecules. It yielded an accuracy of 95% and an AUC of 0.98 in cross validation and of 96% and 0.98% respectively in the case of an independent test set. It was further applied to predict the bitter and sweet taste of natural compounds (70% bitter and 10% sweet with confidence score of 0.60), approved drugs (77% bitter and 2% bitter with a confidence score of 0.75) and acute oral toxicity dataset (75% bitter with a confidence score of 0.75) revealing that toxic compounds are mostly bitter. Moreover, Bayesian based feature

analysis method was applied to discriminate the most occurring chemical features between sweet and bitter compounds. The study revealed that some of the features present in sweet and bitter compounds are not totally independent, while some others tend to be more class specific.

In a recent study, **protein-protein interactions** were predicted by means of random forest framework (Chen and Liu, 2005). Protein interactions are of biological interest because they governate many cellular processes such as metabolic pathways or immunological recognition. Proteins are made of domains and hence their interactions can be considered as domains interactions. In this work, the innovation is that all possible domain interactions are explored, and predictions are based on all protein domains. The sensitivity of the experimental results on Saccharomyces cerevisiae dataset is 79.78%, while specificity is 64.38%. The problem is formulated as a two-class classification problem. The dataset is composed of 4293 unique domains. Hence, each protein pair is represented by a vector of 4293 features where each feature is a domain and has a discrete value (0, 1 or 2). If the sample protein pair does not contain the domain, the feature value is 0, if one of the proteins contain the domain, the feature value is 1 and if both proteins contain the domain, the feature value is 2. This representation is necessary because domains can interact with themselves.

In conclusion, biomedical science has shifted its interest from studying individual molecules to analyzing the interactions of complex molecular and cellular networks that governate biological systems. **Systems biology** is emerging as new subject, with the aim of understanding how the single components of a biological system interact in time and space to determine the system's working in its entirety, and how to treat system's diseases. In this view, AI gives outstanding opportunities for research and tools development.

4.5 Machine Learning in Cardiovascular Risk Prediction

Risk prediction is of one the biggest challenge in clinical cardiology research. Traditional risk models are based on robust regression models. However, these methods are built on a small number of parameters which operate in the same way on everyone and through their range. Therefore, machine learning techniques has been recently introduced in this field, to face challenges that cannot be well addressed by traditional regression methods.

AI and ML can provide a set of instruments to extend and improve the effectiveness of the cardiologist (Johnson *et al.*, 2018). This is a social requirement, for many reasons: the clinical advance in technologies able to produce data such as whole-genome-sequencing will require cardiologists to interpret data from different biomedicine fields; physicians and health care systems are required to be much more efficient and operative today and finally the era of personalized medicine is starting. ML can enhance every step of patient care journey, from research to diagnosis and treatment. See Figure 27 for a summary of the principle areas in which ML will help CV medicine in the future.

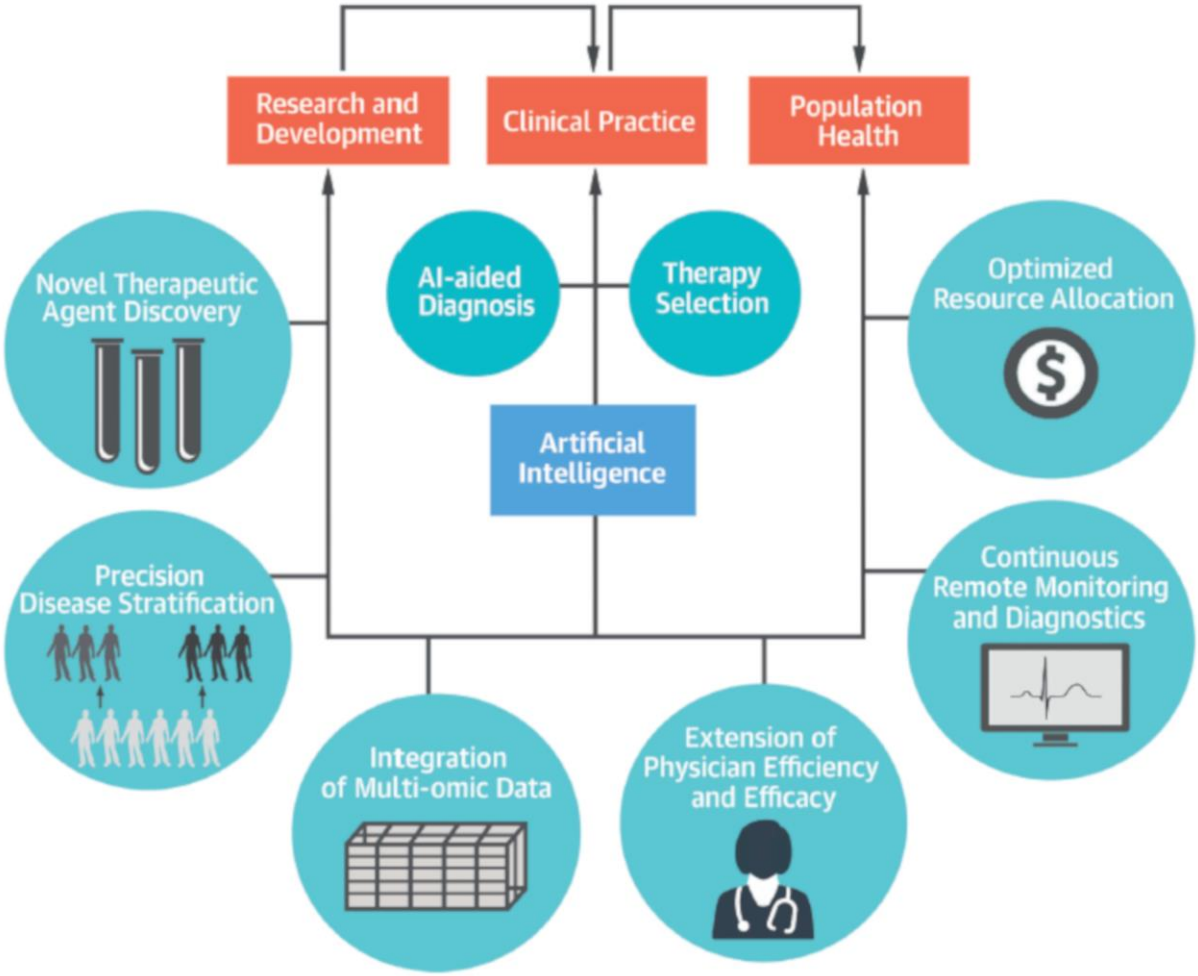

Figure 27 - Illustration of how cardiovascular medicine will be helped by artificial intelligence in the future.

Cardiology needs artificial intelligence because, while with statistical methods some strong assumptions are needed (e.g. independence of observations and no multicollinearity among

variables), ML methods are typically used without making a lot of assumptions about data. Another reason is that when dealing with electronic health record datasets, which are large and messy, ML can help to decide which variable is better to include in the model, by means of feature selection techniques, impossible in the case of statistical regression. Moreover, ML can capture the complex relationships inside data, which are difficult to be seen by traditional statistical methods.

Concerning deep learning in cardiology, it is still developing, and its applications are limited until now. The biggest drawback of deep learning, which prevent it from an intensive use in healthcare, is that it takes an enormous amount of data to be trained and data are often difficult to be acquired in the biomedical field. About unsupervised learning, one of its most promising uses in cardiology is **precision phenotyping** of CV disease. Precision medicine is a contemporary term, describing the synthesis of multiple sources of evidence with the aim of making more precise and personalized diagnosis and treatment. Unsupervised learning enables precision medicine by learning subtypes of diseases.

An explanation of the possible benefits of ML in CV risk prediction is described in the review by Goldstein and others (Goldstein, Navar and Carter, 2016), in which they tried to **predict mortality after diagnosis of acute myocardial infarction** with the aim of explaining traditional methods' limits and how different methods can address different problems. Regression techniques are used most of all in association analyses; however, this is not always the case of prediction analyses, where the attention is pointed at the outcome more than at predictors. Therefore, the constrains used for identifying the effect of the predictors on the output and for adding interpretability for association studies are limitations when these methods are employed in prediction analyses. ML methods, on the contrary, produce a more flexible relationship among variables and outcome. Moreover, they do not require to specify a priori the model structure, but they automatically search for the optimal fit, resulting in a better prediction model, that can however lack in interpretability of predictors-outcome relation. The dataset that was employed in this work is made of 1944 patients that were admitted to hospital with a primary diagnosis of acute myocardial infarction with 43 predictor variables from laboratory tests, demographics and comorbidities. Three challenges of typical regression models that the authors wanted to investigate are:

- Non-linearities: regression models are based on the strict assumption that the risk factors and outcome relations are linear, i.e. the outcome increases uniformly with respect to the

range of a variable. Typical examples of non-linear relations between variables and CV disease outcome are those of age and body mass index, for which CV risk sharply rises together with their increasement.

- Heterogeneity of interactions: it is related to the problem of non-linearity and happens when the relation between a variable and the outcome depends on some other parameter. Typical clinical examples are anthropomorphic characteristics and mortality as well as racial differences in the HDL cholesterol's effects.
- Many predictor variables: with the advent of big data era, datasets are characterized by a large amount of potential predictor variables. It is challenging to understand how many parameters should be used in a risk model and it is important to consider at the same time the number of variables and the number of true events available for each variable. If they are few, estimated effects can be unstable, with high variability. However, also ML methods may become instable in these situations.

There are various ML methods, but each of them is based on a computational algorithm that relates a set of predictors to one or more outcomes. To estimate the model, they randomly or deterministically search for the best fit, trying to balance the trade-off between bias and variance. These two values are both included in a loss function. ML can control this trade-off by tuning algorithms' parameters. The two largest families of ML methods are amendments to regression models and tree-based methods. The most famous examples of the first category are forward and backward selection. They iteratively search for the best subsets of predictors to use and then fit a basic regression model. They are useful when the problem is choosing between many predictors. About the second category, they were born to mimic the way a doctor may approach a patient to make a diagnosis. They are typically able to handle non-linearities, heterogeneous effects and the presence of many variables. One trees' disadvantage is that they present a high variance. However, it is possible to improve this situation by aggregating results from multiple trees (this is the case of random forest). Another approach that does not fit into the two overmentioned groups is KNN. From a medical point of view, it can be the prediction of a patient's outcome based on previous patients with similar symptoms (from the algorithmic point of view, they are represented by a cluster identify by the model). When applying a ML method, parameters must be tuned. A lot of automatic methods exist for this purpose. Then, validation must be performed, which is not necessary in regression models, because one ideally states an analytic model before fitting it to the

data. Conversely, ML methods account for a search for the optimal model. Therefore, validation avoids overfitting. Unlike regression models, ML ones do not have an easily interpretable way to relate predictors and outcome, but they try to summarize the impact of each single parameter into metrics called variable importance. Obviously, ML methods fail when they work with a linear and homogeneous model. In this case, regression models will always perform better. Another area where ML are limited is when causality needs to be included in interpretation. In fact, ML can add a risk predictor into a model only because it improves classification and not because it really causes the outcome. Therefore, results presentation may become harder. ML models cannot be converted into a risk score like, for example, has been done with Framingham risk score.

Table 8 - Models' results in the case of different algorithm (above) and variable importance rankings (below).

	c-Statistic	Squared-error loss	Logistic loss	Misclassification rate[a]
Regression based				
Logistic regression	0.702	0.049	0.995	0.23
Forward selection	0.761	**0.046**	0.995	0.24
LASSO	0.750	**0.046**	0.995	0.26
Ridge	0.753	0.047	0.996	0.27
PCR	0.546	0.049	0.998	0.41
Generalized additive model	0.708	0.050	**0.994**	0.22
Tree based				
CART	0.623	0.053	0.997	**0.12**
Random forests	0.741	0.048	0.995	0.32
Boosting	**0.763**	0.047	0.996	0.20
Other				
Nearest Neighbours	0.583	0.050	0.998	0.22
Neural Networks	0.598	0.065	0.996	0.44

Variable rank	t-Test	GLM	LASSO	GAM	Random forests	Boosting
1	CO_2 Min	Ca^{2+} Max	Ca^{2+} Median	Ca^{2+} Min	CO_2 Min	CO_2 Min
2	CO_2 Median	K^+ Min	Ca^{2+} Max	Ca^{2+} Max	CO_2 Median	WBC Max
3	WBC Max	Hgb Median	Hgb Median	CO_2 Median	WBC Max	CO_2 Median
4	K^+ Max	Ca^{2+} Median	K^+ Median	Ca^{2+} Median	Glucose Max	Ca^{2+} Median
5	CO_2 Max	Hgb Min	K^+ Min	RDW Median	WBC Median	K^+ Max

Results of the current study, in terms of classification results and variables ranking, are shown in Table 8.

Another important current limitation in the application of ML is the need for accurate and reproducible information to build databases. They need to be large and generalizable, to not induce biases in the interpretation or results (Li, Rajagopalan and Clifford, 2014; Obermeyer and Emanuel, 2016). Industry is heavily investing in this field now.

A clinical example is the study performed by Weng and coauthors (Weng *et al.*, 2017) that used tree-based methods and another technique called gradient boosting **to predict CV event risk** in a sample of 378256 British patients. As risk factors variables 22 additional features with potential association with CV disease were included in the analysis, with respect to the eight core baseline variables (gender, age, smoking status, SBP, blood pressure treatment, total cholesterol, HDL cholesterol and diabetes). Four ML algorithms (random forest, logistic regression, gradient boosting machines and neural networks) were used and compared with already existing methods to predict first CV event over 10-years (American College of Cardiology guidelines). Performances were assessed by AUC, sensitivity, specificity, positive predictive value and negative predictive value. They found that, in terms of AUC, random forest improved of 1.7%, logistic regression of 3.2%, gradient boosting of 3.3% and neural network of 3.6%, outperforming the American College of Cardiology and American Heart Association risk algorithm (AUC 0.728, 95% CI 0.723-0.735). The best ML algorithm (neural network) correctly predicted +7.6% patients who developed CV diseases, with respect to traditional algorithm. Figure 28 shows the top 10 risk factors identified by each algorithm. Interestingly, while several traditional risk factors are present as top ranked ones for all ML methods, diabetes was not.

ACC/AHA Algorithm		Machine-learning Algorithms			
Men	**Women**	**ML: Logistic Regression**	**ML: Random Forest**	**ML: Gradient Boosting Machines**	**ML: Neural Networks**
Age	Age	Ethnicity	Age	Age	Atrial Fibrillation
Total Cholesterol	HDL Cholesterol	Age	Gender	Gender	Ethnicity
HDL Cholesterol	Total Cholesterol	SES: Townsend Deprivation Index	Ethnicity	Ethnicity	Oral Corticosteroid Prescribed
Smoking	Smoking	Gender	Smoking	Smoking	Age
Age x Total Cholesterol	Age x *HDL Cholesterol*	Smoking	*HDL cholesterol*	*HDL cholesterol*	Severe Mental Illness
Treated Systolic Blood Pressure	Age x Total Cholesterol	Atrial Fibrillation	HbA1c	Triglycerides	SES: Townsend Deprivation Index
Age x Smoking	Treated Systolic Blood Pressure	Chronic Kidney Disease	Triglycerides	Total Cholesterol	Chronic Kidney Disease
Age x *HDL Cholesterol*	Untreated Systolic Blood Pressure	Rheumatoid Arthritis	SES: Townsend Deprivation Index	HbA1c	*BMI missing*
Untreated Systolic Blood Pressure	Age x Smoking	Family history of premature CHD	BMI	Systolic Blood Pressure	Smoking
Diabetes	Diabetes	COPD	Total Cholesterol	SES: Townsend Deprivation Index	Gender

Figure 28 - Variables ranking with respect to their importance determined by the coefficient effect size for the traditional method and ML methods.

Another clinical example employing support vector machine is the one by Cui et at. (Cui *et al.*, 2017). Clinicians may find SVMs useful, because they are relatively simple and can capture complex non-linear relationships. In this work, it has been demonstrated the usefulness of SVMs

for **prediction of in-stent restenosis from plasma metabolite levels**, with 90% accuracy. In-stent restenosis is a big issue for patients who have undergone percutaneous coronary intervention, therefore the identification of new biomarkers to predict it can be very important for patient care. The biggest disadvantages of SVM classification are the fact that they do not perform probabilistic classification, but they work by default on dichotomized outcomes and the fact that computation of input variables in a very high-dimensional space can be difficult or impossible. In this study, plasma metabolomic biomarkers were evaluated as diagnostic tools. 400 patients were used in the discovery step, while other 500 in the validation phase. Results show that a set of 6 plasma metabolites belonging to sphingolipid and phospholipid metabolism can predict with 91% sensitivity and 90% specificity in-stent restenosis during the learning phase, while with 90% accuracy (95% CI, 87% to 100%) during the validation phase. Sets of 5 to 17 metabolites were manually chosen from the top 58 to be used in classification, with two multivariate methods: random forest and linear support vector machine. Classification performance was evaluated by AUC. This is the demonstration of the powerful predictive value of plasma biomarkers when compared to traditional imaging techniques. Moreover, patients at risk can be identified early and consequently treatment strategies can be changed on time.

Another example of ML applied to cardiovascular risk prediction is represented by a **comparison of ML techniques in the domain of heart disease** (Pouriyeh *et al.*, 2017). The dataset used is the Cleveland Heart Disease dataset, freely available at the University of California Irvine (UCI) machine learning repository (https://archive.ics.uci.edu/ml/index.php). It contains 303 instances and 75 attributes, but every referring publication used only 14 of them, because they are closely linked to heart disease. They are age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercised induced angina, oldpeak, slope, number of vessels colored and thalassemia. Then there is the predicted attribute, reporting the presence or absence of heart disease in the patient. It is an integer ranging from 0 (no presence) to 4. Every study on the Cleveland database has considered only two possible classes (0, meaning absence of heart disease or 1, 2, 3, 4, meaning presence). It has already been used by many researchers to try different classification algorithms and machine learning techniques on heart disease data. Among the most relevant publication, Detrano performed a logistic regression algorithm with a 77% classification accuracy (Detrano *et al.*, 1989), Gudadhe et al. combined the multilayer perceptron network with the support vector machine approach with a 80.41%

classification accuracy (Gudadhe, Wankhade and Dongre, 2010), Kahramanli and Allahverdi used a fuzzy neural network in combination with an artificial neural network with a 87.4% classification accuracy (Kaharamanli and Allahverdi, 2008) and Palaniappan and Awang developed an intelligent heart disease prediction system, made of a combination of different data mining techniques with different classification accuracies (Palaniappan and Awang, 2008). In this paper, decision tree, naïve Bayes, multilayer perceptron, k-nearest neighbor, single conjunctive rule learner, radial basis function and support vector machine are used. Moreover, the ensemble prediction of classifiers, including bagging, boosting and stacking are employed. Support vector machine outperformed all other methods in terms of accuracy (84.15%). Table 9 reports results of each single classifier.

Table 9 - Standard metrics for all employed classifiers.

Classifier	Precision	Recall	F-Measure	ROC Area	Accuracy (%)
Decision Tree(DT)	0.774	0.830	0.801	0.800	77.55
Naïve Bayes (NB)	0.836	0.867	0.851	0.904	83.49
K Nearest Neighbor (K-NN, K=1)	0.782	0.782	0.782	0.752	76.23
K Nearest Neighbor (K-NN, K=3)	0.821	0.836	0.829	0.838	81.18
K Nearest Neighbor (K-NN, K=9)	0.848	0.842	0.845	0.898	83.16
K Nearest Neighbor (K-NN, K=15)	0.847	0.836	0.841	0.904	82.83
MultiLayer Perceptron (MLP)	0.824	0.824	0.824	0.894	82.83
Radial Basis Function (RBF)	0.845	0.861	0.853	0.892	83.82
Single Conjunctive Rule Learner (SCRL)	0.734	0.703	0.718	0.707	69.96
Support Vector Machine (SVM)	0.827	0.897	0.860	0.836	84.15

A comparative study between Framingham and quantum neural network based approach in the field of CV risk prediction has been done (Narain, Saxena and Goyal, 2016). Data from 689 patients showing CVD symptoms were used together with data from 5209 Framingham study patients, for validation purposes. This system achieved 98.57% accuracy, significantly higher when compared with FRS. Quantum neural networks are based on multi-level transfer function. Differently from artificial neural networks, their hidden units use a nonlinear activation function (consisting of linear superposition of multi-sigmoid functions). This paper demonstrates how FRS (it was developed in the 1960s) is out of date and new parameters should be considered. Moreover,

the major limits of FRS are its ineffectiveness with respect to value ranges, in fact it is not applicable to patients aged 20 to 100 years and the maximum threshold limit is 30%.

ML has also on its way to echocardiography imaging, the so-called heart of cardiology (Tajik, 2016). It produces digital images of high spatial and temporal resolution and their analysis can be automatized (in terms of chamber dimensions, volumes, wall thickness and motion) by means of intelligent systems. They can drastically reduce the current time needed for echocardiographic examination. The next phase will be an automatically interpretation of this kind of images (reducing intraobserver and interobserver variability as well as cognitive errors). An interesting study is the one in which ML has been employed **to automate morphological and functional evaluations in 2D echocardiography** (Narula *et al.*, 2016). ML models can aid cardiac phenotypic recognition, working on features of cardiac tissue deformation: this is the base of this work, which tried to understand if a ML framework, incorporating echocardiographic data, can have a diagnostic value in the discrimination of hypertrophic cardiomyopathy from physiological hypertrophy in athletes. Dataset was composed by 77 healthy people and 62 with the disease. An ensemble method was employed (made by a combination of support vector machines, random forests and artificial neural networks), based on majority voting for prediction and on cross validation. By integrating the results of three different classifiers, additional assurance of validity can be guaranteed even if with small databases. Variables' importance was also evaluated by means of information gain criterion and Ranker method. Volume was identified as the best predictor. Results suggest that ML can help in this type of diagnosis and to develop a real-time system for automated interpretation of clinical situations. In fact, the model obtained 96% sensitivity and 77% specificity for differentiating the normal from the pathological condition, showing equal sensitivity, but improved specificity when compared to conventional methods.

5. Cardiovascular Risk Prediction in Rheumatic Patients by Artificial Intelligence Paradigms

5.1 Introduction

According to World Health Organization, cardiovascular diseases (CVDs) are the first cause of death globally. Most CVDs can be prevented by pointing at behavioral risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol through population-wide strategies. The early detection and the preventive treatment of CVDs are crucial for people with high CV risk and heart monitoring is decisive. Therefore, in the last years, heart monitoring systems have growing, with special attention to systems that produce local online real-time classification, such as mobile monitoring (Sannino and De Pietro, 2011). However, CVDs diagnosis is very tricky even for experts, due to the presence of many concurrent risk factors, some of which uncertain or unknown, and consequently may be incorrect or delayed. To help physicians avoiding errors during diagnosis, many researchers have tried to apply machine learning techniques to heart disease (Detrano *et al.*, 1989; Kaharamanli and Allahverdi, 2008; Palaniappan and Awang, 2008; Gudadhe, Wankhade and Dongre, 2010).

In patients affected by chronic inflammatory arthritis (such as rheumatoid arthritis, psoriatic arthritis and ankylosing spondylitis) an increased CV risk has been observed. The augmented presence of traditional CV risk factors, chronic inflammation and potential adverse effects of drugs (such as glucocorticoids or nonsteroidal anti-inflammatory drugs) contribute to affect and worsen rheumatic patients' lives. Traditional CV risk calculators (such as Framingham, CUORE and SCORE risk score) present limitations in the prediction of CVDs in patients with inflammatory arthritis. They tend to underestimate CV risk and the different scores are poorly calibrated for rheumatic patients (Arts *et al.*, 2015; Navarini *et al.*, 2018), even when corrections to algorithms have been considered (Agca *et al.*, 2017). Moreover, standard methods make the general assumption that each risk factor is linearly related to CVD outcome. Hence, these methods could not capture the increased complexity that characterizes rheumatic patients and oversimplify complex variables' relations, most of all in situations with a big number of non-linearly related variable. Different solutions may be adopted: redefining the way by which already available algorithms differentiate low risk patients from high risk ones and developing more specific CV risk

algorithms or with more biomarkers and disease-related CV risk factors. Surely, prospective and larger studies to improve CV risk prediction in patients affected by inflammatory arthritis are needed. Another possibility is to employ machine learning (ML) techniques to predict CV risk in rheumatic patients. ML was recently introduced in cardiology to face challenges that cannot be solved by traditional statistical methods (Johnson *et al.*, 2018). For example, a comparison of ML techniques in the domain of heart disease was performed by Pouriyeh and coworkers (Pouriyeh *et al.*, 2017). A comparative study between Framingham and quantum neural network based approach (Narain, Saxena and Goyal, 2016) showed how Framingham is out of date and the outstanding potential of ML applied to CV risk prediction. Hence, in this work we explore the application of ML in the specific field of CV risk prediction in rheumatic patients.

ML is a subfield of artificial intelligence that can be described as the ability of computers to learn how to solve a given problem without being explicitly programmed for this. The learning process is made nowadays possible by deriving knowledge from the huge quantity of data present in almost every field (the so-called big data) and has the objective of making predictions. The two biggest subsets of ML are supervised learning and unsupervised learning. In the first case, the model is built from a database that already contains the desired output, such as CVD outcome. In the second case, there is no prior knowledge about the event inside the dataset, therefore the model aims at finding subgroups of the original dataset which have common features. ML does not present the same limitations as in the case of traditional statistical methods. In particular, not many assumptions must be made on the underlying data and non-linearities can be catch more easily. Also, ML can identify hidden variables of a model, by inferring them from other variables.

In this study, ML techniques were used with the aim of elucidate unknown complex relationships driving patients with inflammatory arthritis (with attention to patients with psoriatic arthritis, ankylosing spondylitis and systemic lupus erythematosus) towards an increased CV risk. This is an innovative approach, because ML has never been applied before to CV risk prediction in rheumatic patients. The research is a first attempt towards the development of newly efficient, personalized and reliable CV predictors to be used in clinics. First, a comparison between traditional (Framingham, CUORE and SCORE risk scores) and novel techniques (support vector machine, random forest and k-nearest neighbor) has been performed, to explore performances of traditional risk predictors on general population and rheumatic patients, with and without European League Against Rheumatism (EULAR) correction coefficient (Agca *et al.*, 2017). It is a

multiplicative factor of 1.5 (to be applied on the final CV risk score results) proposed by the European association to account for the increased CV risk in rheumatic patients. Second, ML techniques (support vector machine, random forest and k-nearest neighbors) have been applied to the problem as a new CV risk model. In addition to this, feature analysis (by means of random forest's importances) has been performed, to tackle a higher number of variables than those used in traditional risk predictors. The above-mentioned analysis is relevant to understand the possible key players in cardiovascular risk among typical rheumatic patients' characteristics. These novel techniques can help in personalizing the risk prediction towards a specific need (such as patients with psoriatic arthritis).

5.2 Materials and Methods

The activity flow of the work is divided in the following phases:

1. database definition,
2. algorithms selection and development,
3. dataset preprocessing and features analysis,
4. classifiers training and validation,
5. classifiers evaluation and features importance.

The aim of this study was to make predictions about cardiovascular risk in patients with inflammatory arthritis (PsA, AS and SLE), therefore supervised learning approach was used.

5.2.1 Database Definition

In the current work, four databases were employed:

1. 3658 American patients, from the Framingham heart study, retrieved from Kaggle website (https://www.kaggle.com/datasets). Risk factors included in this dataset are: gender (0: female, 1: male), age (years), smoking status (0: nonsmoker, 1: smoker), hypertension treatment (0: not treated, 1: treated), total cholesterol (mg/dl), systolic blood pressure (SBP, mmHg), body mass index (BMI, kg/m^2), diabetes (0: without diabetes, 1: with diabetes) and CVD event (0: without CVD, 1: with CVD). In this dataset, 557 patients had a CVD and 3101 not. This dataset is considered representative for a general population. From now on, this dataset will be indicated as the **general dataset**.

2. 155 Italian patients with psoriatic arthritis (PsA), provided by Prof. Afeltra group. Risk factors included in this dataset are the same included in the general dataset with the addition of: pathology time window (PTW, years), CVD family history (0: no, 1: yes), atrial fibrillation (AF, 0: no, 1: yes), HDL cholesterol (mg/dl), weight (kg), height (cm), c-reactive protein (CRP, mg/l), axial arthritis (0: no, 1: yes), peripheral arthritis (0: no, 1: yes), enthesitis (0: no, 1: yes), dactylitis (0: no, 1: yes), psoriasis (0: no, 1: yes), psoriasis area severity index (PASI, number), onychopathy (0: no, 1: yes), inflammatory bowel disease (IBD, 0: no, 1: yes), uveitis (0: no, 1: yes), comorbidity (0: no, 1: yes), use of statins (0: no, 1: yes) and CVD event. In this dataset, 21 patients had a CVD and 134 not in the case of Framingham assumptions, 15 yes and 140 not in the case of CUORE assumptions and 17 yes and 138 not in the case of SCORE assumptions. From now on, this dataset will be indicated as the **PsA dataset**.
3. 133 Italian patients with ankylosing spondylitis (AS), provided by Prof. Afeltra group. Risk factors included in this dataset are the same included in the general dataset with the addition of: PTW, CVD family history, AF, HDL cholesterol, use of cardio aspirin (0: no, 1: yes), CRP, peripheral arthritis, enthesitis, dactylitis, IBD, uveitis, diabetes, comorbidity, use of statins and CVD event. In this dataset, 18 patients had a CVD and 115 not. From now on, this dataset will be indicated as the **AS dataset**.
4. 194 Italian patients with systemic lupus erythematosus (SLE), provided by Prof. Afeltra group. Risk factors included in this dataset are the same included in the general dataset with the addition of: PTW, CVD family history, AF, HDL cholesterol, weight, height, BMI, IBD, uveitis, diabetes, use of statins, lupus nephritis (0: no, 1: yes), chronic kidney disease (CKD, 0: no, 1: yes), metabolic syndrome (MeS, 0: no, 1: yes), average monthly dose of glucocorticoids in mg equivalent of prednisone (GCdose, number), use of aspirin (ASA, 0: no, 1: yes), synthetic antimalarials treatment (HCQ, 0: no, 1: yes), number of anti-phospholipids positivity (aPL number, 0: no, 1: yes), organ disease index (SDI, number), disease activity index (SLEDAI, number), year disease reactivations number before follow-up (FLARES, number) and CVD event. In this dataset, 21 patients had a CVD and 134 not. From now on, this dataset will be indicated as the **SLE dataset**.

It is worth noticing that, in each database, about 15% of patients had a CVD event, therefore this is a case of classes' imbalance.

5.2.2 Algorithms Selection and Development

All algorithms were developed in Python 3.7.2, with the help of the following scientific computation libraries: NumPy, to manipulate data; Scikit-learn, to implement ML pipelines; Pandas, to manipulate data at a higher level than with NumPy and Matplotlib, to visualize data.

First, traditional risk predictions (Framingham, CUORE and SCORE risk scores) were implemented and calculated for every database, using the traditional formula and the one corrected by EULAR coefficient. Performance metrics were calculated for the two cut-offs: low-to-intermediate (10% in the case of Framingham and CUORE, 1% in the case of SCORE) and intermediate-to-high (20% in the case of Framingham and CUORE, 5% in the case of SCORE). Then, ML techniques were employed: support vector machine (SVM), random forest (RF) and k-nearest neighbor (KNN).

5.2.3 Dataset Preprocessing and Features Analysis

Data were always standardized to take them on the same scale. ML techniques were first applied to the general dataset, using only Framingham features (i.e. age, sex, SBP, total cholesterol, smoking status and hypertension treatment). The same traditional features were used also on rheumatic datasets, to make results comparable with the traditional ones. The general dataset does not contain missing values, while the PsA, AS and SLE datasets contain several missing values in rheumatic features. Therefore, features with more than 40 missing values were removed, while the others were imputed, because by removing them there was the possibility to remove important information from the dataset or, for small dataset like the one we have, to compromise model's reliability. 40 seems a reasonable assumption with respect to the total number of patients for each dataset (PsA dataset: 155, AS dataset: 133, SLE dataset: 194), because in each case only about one third of the data is imputed. Imputation followed this strategy: in binary attributes missing values were substituted by 0 (absence of event), while in numeric attribute they were substituted by the normal value over the general Italian population (data taken from literature).

5.2.4 Classifiers Training and Validation

To train the classifiers, a balanced dataset was used (i.e. equal number of patients with and without CVD event, about 600 samples). Then, the classifier was tested on an unbalanced dataset (about 15% of patients which had a CVD event), composed by the remaining data not used during training (about 3100 samples). Bootstrap technique was used to assess model performance, with 25 random

splits with replacement. It is a sampling technique which, by means of iterative dataset's random splits, gives the possibility to calculate different times algorithms performances on different patients' subsets, making performance evaluation more reliable. A graphical explanation of the dataset partition is shown in Figure 29.

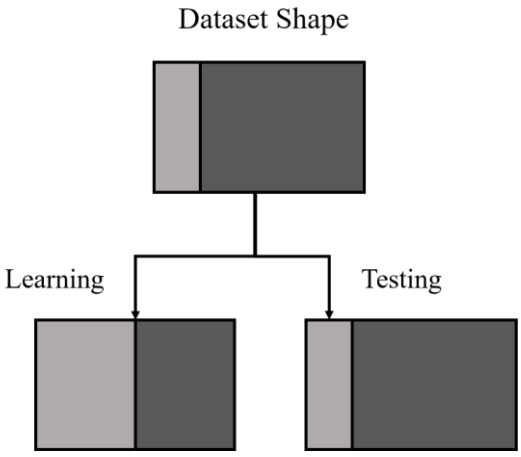

Figure 29 - Dataset partition for machine learning models' training and testing. The dataset was originally made of about 15% patients who experienced a CV event (light grey bar).

Models' hyperparameters (SVM: C in the case of a linear kernel and C and γ in the case of a radial basis function kernel; RF: number of trees and splitting criterion; KNN: number of neighbors, i.e. k) were optimized by means of grid search, setting AUC as scoring function and performing a 4-fold cross validation. ML classifiers presented the following optimized parameters after grid search:

- SVM: radial basis function kernel, C = 0.1 and γ = 0.01;
- RF: entropy as splitting criterion and 500 trees;
- KNN: minkowski distance metrics and K = 25.

Training was not necessary in the case of traditional statistical methods, because they have already been implemented in previous studies and regression coefficients have already been estimated.

5.2.5 Classifiers Evaluation and Features Importance

Discriminatory ability for the algorithms was assessed by ROC curves and AUC values, sensitivity, specificity, accuracy and odds ratio (OR). Calibration between predicted and observed events was evaluated by Hosmer-Lemeshow tests, by comparing the agreement of CV events in groups

stratified in deciles. First, traditional algorithms (FRS, CUORE and SCORE) were evaluated on the general population and on the rheumatic one. Second, ML techniques were applied the general population, by means of bootstrap technique and performances were compared with FRS as reference for traditional methods. Finally, obtained models trained on the general population were validated on the PsA, AS and SLE datasets.

Feature importance analysis was performed on PsA, AS and SLE datasets through importances of RF, which was pre-trained on balanced datasets using all rheumatic features. This step had the aim of evaluating each variable's role and importance as CV risk predictive parameters.

5.3 Results

Sensitivity, specificity and accuracy values are reported in the case of the general and PsA dataset for the two cut-off values in Figure 30. Looking at the general dataset plot, it can be seen the trade-off of choosing the right cut-off. In fact, lowering the cut-off value (e.g. from 20% to 10%) causes an increase in sensitivity, but also a decrease in specificity. The accuracy remains similar in the two cases. Therefore, finding the right cut-off for a specific kind of patients is tricky. Moreover, CUORE performances are worse than the others. A possible reason can be the fact that the general dataset is composed by American patients, while CUORE was developed using an Italian cohort. Concerning PsA dataset, Framingham sensitivity drastically decreases, hence it is clear that traditional algorithms underestimate CV risk in rheumatic patients. Numerical performances values are reported in Table S10 and Table S11 of supporting information section.

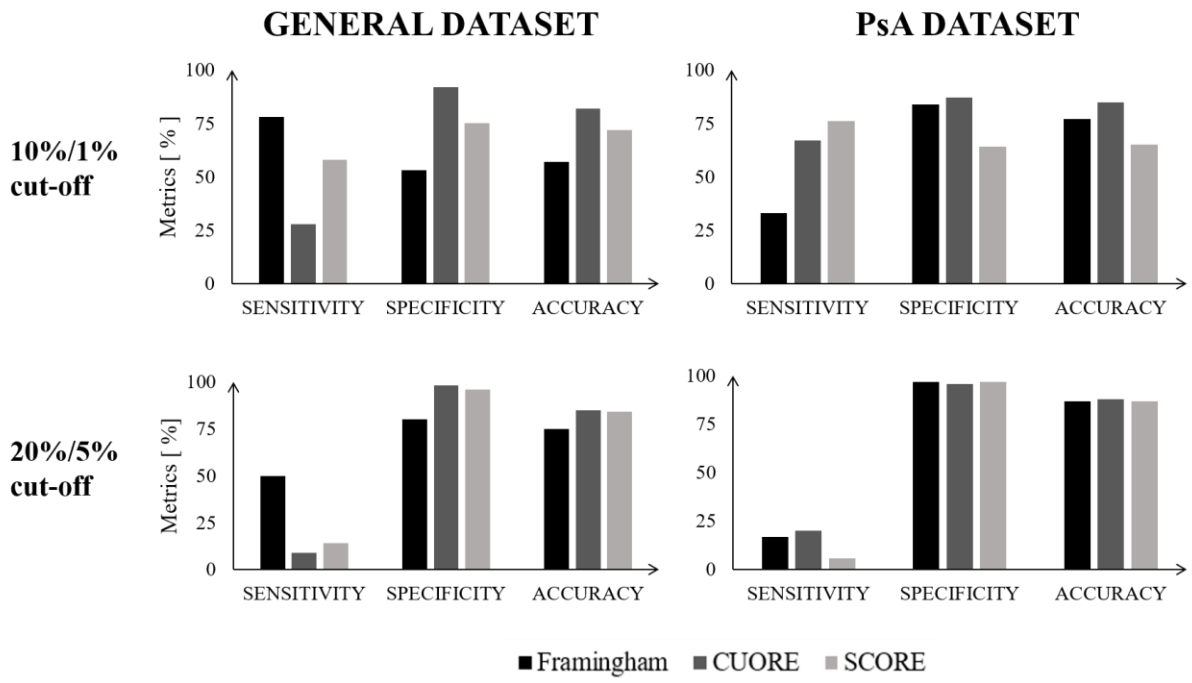

Figure 30 - Traditional risk algorithms performance. Framigham results are represented in black, CUORE ones in grey and SCORE ones in light grey.

Another proof of traditional risk algorithms' lower performance on rheumatic patients (with respect to general patients) is highlighted by observed versus predicted CV events in deciles of predicted risk for Framingham, when compared with the same plot realized applying Framingham on the general dataset, shown in Figure 31. Framingham risk score appears poorly calibrated for PsA patients and the risk observed exceeds that predicted. CUORE and SCORE perform in a similar way.

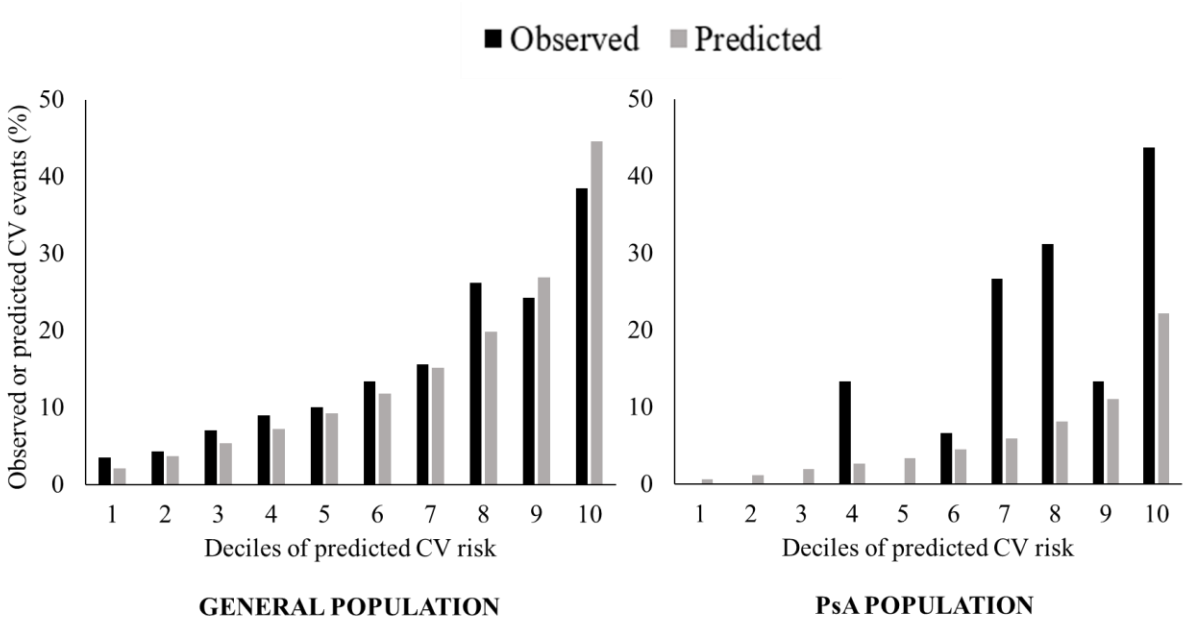

Figure 31 - Observed versus predicted CV events (%) in deciles of predicted risk for Framingham in the case of the general dataset and PsA one. Observed events are represented in black, while predicted events in grey.

Concerning EULAR correction factor, values in the case of Framingham dataset are reported in Figure 32. Framingham, CUORE and SCORE algorithms perform in a similar way looking at all metrics, when EULAR correction factor is applied. As shown in Figure 32, EULAR coefficient increases performance in sensitivity, lowering the specificity. Therefore, EULAR correction factor affect the metrics similarly to the cut-off strategy. Moreover, the discriminative ability and calibration are still limited.

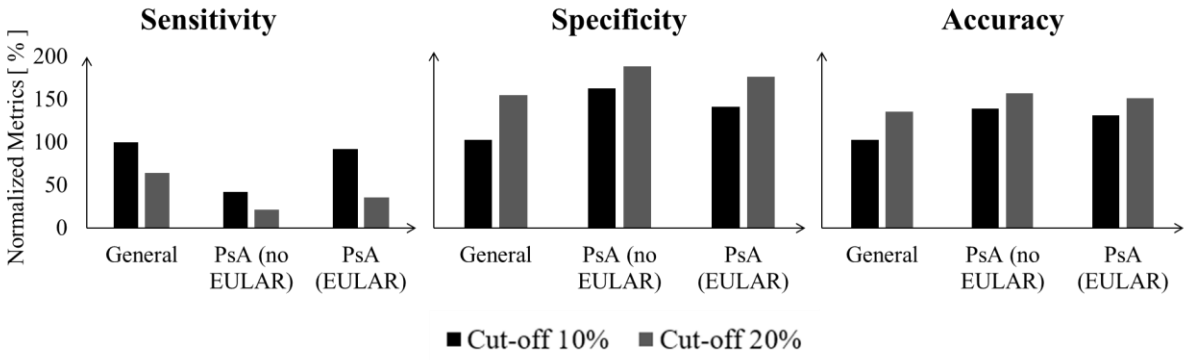

Figure 32 - EULAR correction factor applied to Framingham risk score, in the case of the general dataset and the PsA one. Data are normalized with respect to the general case at 10% cut-off. Black bars represent data at 10% cut-off, while grey bars at 20% cut-off.

ML techniques (SVM, RF, KNN) were firstly applied to the general dataset. Results in terms of performances are shown in Figure 33. ML shows overall comparable results when compared with traditional algorithms (in this case, Framingham 10% cut-off with EULAR correction factor was chosen as reference for comparison). Interestingly, ML does not consider any cut-offs and maintains a higher specificity with similar sensitivity, when compared with FRS.

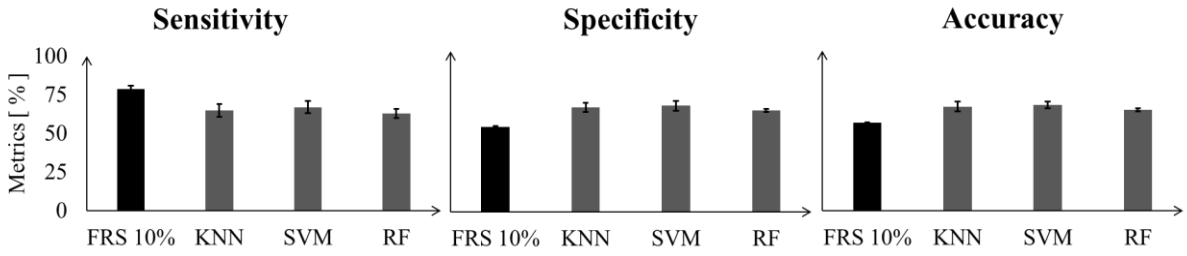

Figure 33 - Performance metrics calculated in the case of general population, comparing FRS 10% cut-off with machine learning algorithms (SVM, RF and KNN) on the same test set. Standard deviations are represented as black bars. FRS results are shown in black, while ML results in grey. Bootstrap technique was employed.

Then, ML techniques were applied on rheumatic patients. Models were built on the general population and validated over the rheumatic one. Results for PsA, AS and SLE datasets are shown in Figure 34, which represents a comparison between Framingham (10% cut-off EULAR version) performances and ML algorithms. ML is trained only with 6 Framingham features (gender, age, SBP, hypertension treatment, smoking status and total cholesterol) and exhibits higher sensitivity in PsA, whereas show in general lower performances in AS and SLE. AUC values in the case of

PsA population are: 0.7747 (95% CI 0.674-0.865) for FRS, 0.8468 (95% CI 0.745-0.925) for SVM, 0.8452 (95% CI 0.746-0.937) for RF and 0.8010 (95% CI 0.705-0.881) for KNN. AUC values in the case of AS population are: 0.6489 (95% CI 0.443-0.827) for FRS, 0.6990 (95% CI 0.545-0.849) for SVM, 0.7297 (95% CI 0.605-0.845) for RF and 0.8010 (95% CI 0.496-0.770) for KNN. AUC values in the case of SLE population are: 0.6775 (95% CI 0.528-0.815) for FRS, 0.6909 (95% CI 0.567-0.809) for SVM, 0.6839 (95% CI 0.555-0.796) for RF and 0.7594 (95% CI 0.668-0.853) for KNN. Numerical performances values are reported in Table S12 of supporting information section.

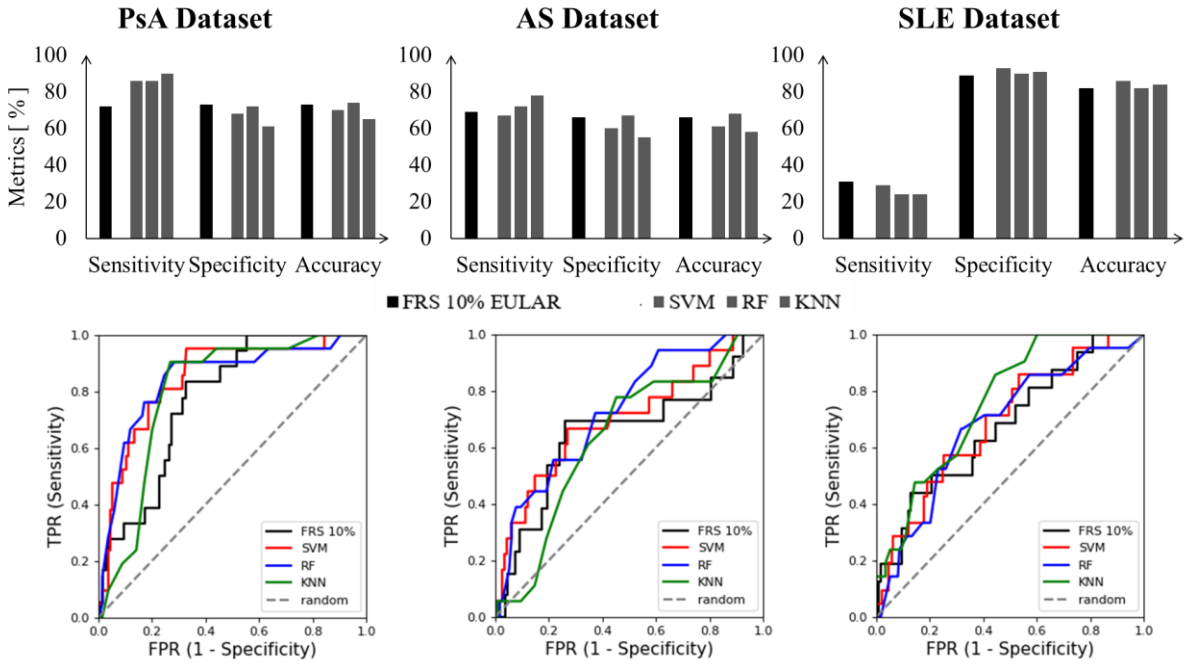

Figure 34 - Performance metrics in the case of validation of machine learning models (SVM, RF, KNN) on PsA, AS and SLE populations (all models were trained on the general population with 6 features: sex, age, smoking status, total cholesterol, systolic blood pressure, hypertension treatment).

Observed versus predicted CV events in deciles of predicted risk for KNN, SVM and RF algorithms respectively are shown in Figure 35. All three algorithms appear poorly calibrated for PsA patients, but better calibrated than FRS. They overestimate CV risk. Probability scores are better distributed then in the case of FRS and, especially with SVM and RF, observed risk presents a more regular shape.

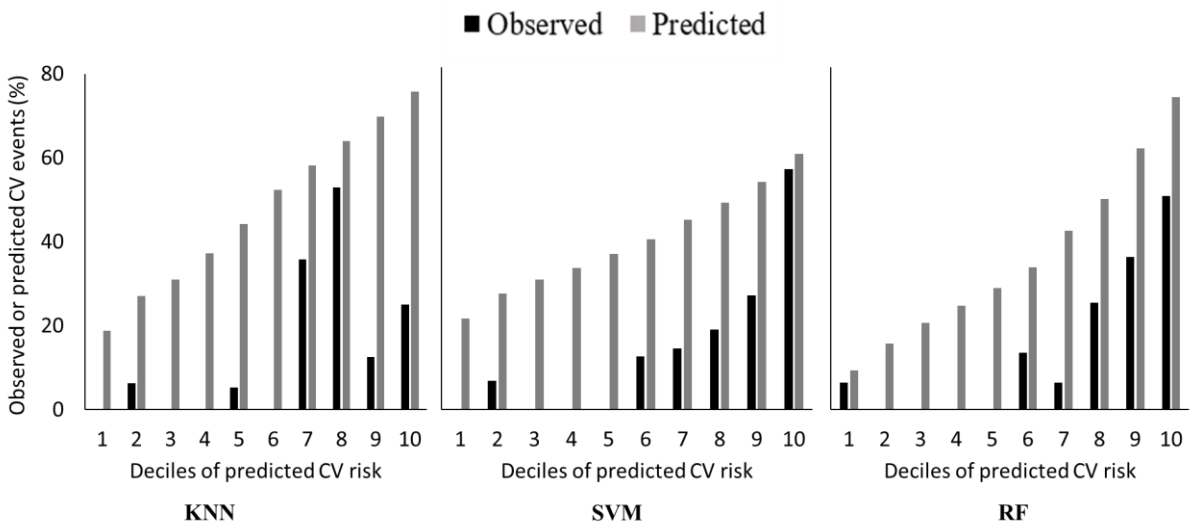

Figure 35 - Observed versus predicted CV events (%) in deciles of predicted risk for KNN, SVM and RF in the case of PsA patients.

Feature analysis was also performed in this study, by means of RF's importances. RF was pre-trained in PsA, AS and SLE datasets using all rheumatic features and it ranked features based on their relative importance. Results are represented in Figure 36. It is evident from the plot that in the case of PsA most of traditional Framingham features are among the best features because they cover about 50% of classification weight, while for AS and SLE is not the same. In the case of AS, CRP has the highest importance, while SBP and hypertension treatment has lower importance with respect to PsA case. In the case of SLE the situation is even worse than AS, because the best features are represented almost only by typical rheumatic features. This is coherent with the previous result, in which ML model had better performances than Framingham only on PsA dataset. The reason is that the model was developed using only 6 traditional Framingham features and therefore feature importance analysis might be crucial to select variables to be included in further risk predictors development.

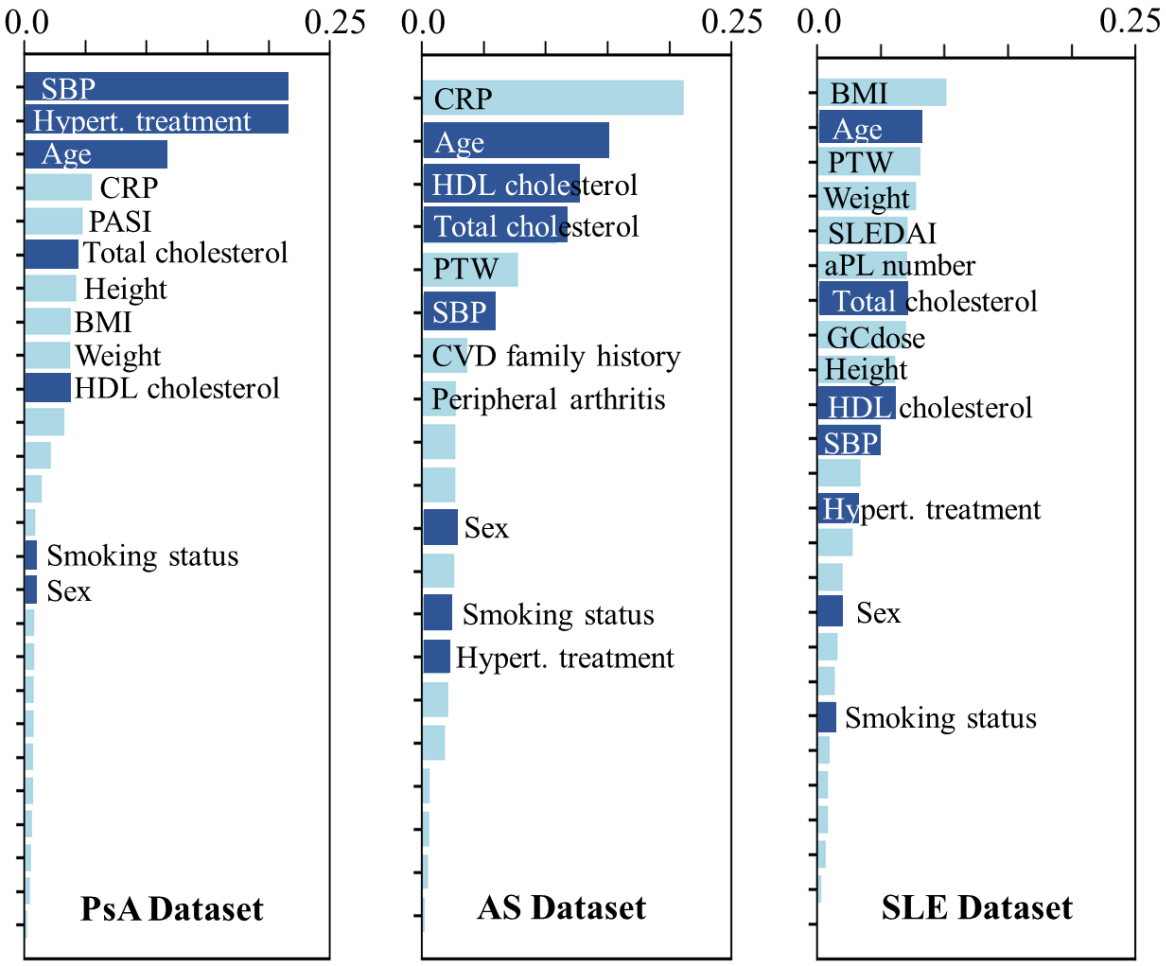

Figure 36 - Features analysis by means of random forest's importances in the case of PsA, AS and SLE populations. The dataset used for RF pre-training has an equally number of patients without a cardiovascular event and with cardiovascular event. In dark blue, traditional Framingham study features are highlighted, while the ones in light blue are typical rheumatic features.

5.4 Discussion

In the case of the general population, we obtain a sensitivity of 78.3% with 10% FRS and 49.6% with 20% FRS and a specificity of 52.8% and 80.0% respectively for the two cut-offs. Artigao-Rodenas and coworkers obtained, in a southern Europe population, 75% specificity with a cut-off of 26.6% for men and of 14.2% for women; sensitivity was respectively 74% for men and 61.3% for women (Artigao-Rodenas *et al.*, 2013). In another work, when comparing FRS and SCORE algorithms, a sensitivity of 57.7% and a specificity of 82.6% were obtained for the overall

population in the case of FRS with a 20% cut-off value (Gunaydin *et al.*, 2015). These results are comparable. CUORE and SCORE perform in a similar way to FRS. From this first part of the study it is clear the cut-offs effect on these scores: when growing the cut-off, an increasement in sensitivity occurs with a correspondent reduction of specificity. Accuracy remains the same or slightly decreases. EULAR correction factor acts in a similar way to the cut-off strategy. In the case of rheumatic patients, PsA patients have 33.3% sensitivity and 83.6% specificity with FRS 10% cut-off and 16.7% sensitivity and 96.9% specificity with FRS 20% cut-off. Therefore, sensitivity drastically decreases when predicting CV risk in PsA patients and EULAR coefficient does not present an acceptable improvement. Traditional models show also a poor calibration on PsA patients as we can see in Figure 31. They underestimate the risk in this kind of population (Navarini *et al.*, 2018).

For this reason, we explored the application of ML methods as new CV risk models for rheumatic patients. The general dataset has been used to evaluate ML performances and to develop stable models, thanks to the large number of patients which it contained. Better results were obtained using a balanced dataset (i.e. with the same number of patients who experienced a CV event and patients who do not) to learn the ML algorithms.: this is obvious, because ML methods learn from data, therefore if the input data is asymmetric, the model will learn a consequent asymmetric decision rule. This concept opens interesting possibilities in the tuning of the algorithms. For example, if identifying patients who are healthy is more important that identifying patients who are not, an unbalanced dataset with more healthy patients can be used to learn a system. ML models are more stable than traditional methods, as it can be seen in Figure 33, with equivalent low standard deviations. Parameters' optimization has been always performed only on training dataset, otherwise performances would have been too optimistic. Best classification results were obtained in the case of PsA patients, when applying to them the ML models trained on the general dataset. SVM has 85.7% sensitivity, 67.9% specificity and 70.3% accuracy; RF 85.7% sensitivity, 72.4% specificity and 74.2% accuracy and KNN 90.5% sensitivity, 67.9% specificity and 70.3% accuracy. These results outperformed sensitivity with respect to FRS, but they tend to overestimate the risk, therefore future work is necessary to overcome this limitation. Calibration plots show that these models are better calibrated than FRS on this kind of patients, even if they have the opposite problem with respect to FRS: while it underestimates CV risk, they overestimate it (Figure 35). In a study about the employment of statistics and deep belief networks to CV risk prediction (Kim,

Kang and Lee, 2017), deep belief networks performed better than other prediction methods using six variables (age, SBP, diastolic blood pressure, HDL cholesterol, smoking status and diabetes). SVM had 100% specificity, 71.8% sensitivity and 71.8% accuracy, hence it was effective in identifying low risk, but it could not correctly predict high risk. RF had 61.4% specificity, 82.2% sensitivity and 77.2% accuracy. Statistical deep belief networks outperformed all methods, with 73.3% specificity, 87.6% sensitivity and 83.9% accuracy. However, a better results was obtained by Narain and coworkers, who did a comparison between FRS and quantum neural network based approach (Narain, Saxena and Goyal, 2016), with 98.57% accuracy. This result shows that different ML methods could be used in CV risk prediction and in the specific field of rheumatic patients, with big potentialities.

Even if PsA, AS and SLE are all rheumatic pathologies, it is clear from the variables' importance that they present different characteristics and therefore different descriptive features (Cooksey *et al.*, 2018). Therefore, future models should consider this complexity. RF's importances is a useful technique to understand the variables that are the most informative inside a bigger set. However, this method does not consider the possible correlation between two variables. Hence, results must be carefully analyzed, because it is possible that, if two variables are highly correlated, one will be placed between the most important variables and the other one between the less relevant ones. A possible explanation to the fact that ML gave worse results in the case of AS and SLE populations is that they do not present among the top features the traditional ones used by FRS. In AS patients, the most important variable is CRP. This is a result that was already clinically validated, as in the study performed by Benhamou and coworkers (Benhamou, Gossec and Dougados, 2010).

In this study, good results were obtained using only 6 traditional features. This is a promising result, because in the future, adding more specific biomarkers could improve the classification performances and the output model would not be too complex or difficult to understand.

5.5 Conclusions and Future Developments

CVDs remain a big health issue, responsible for 3 to 9 million deaths every year only in Europe. Rheumatic patients are at high CV risk and need therefore preventive treatment to better their life conditions. The biggest problem with existing prediction models is that they typically underestimate the CV risk in this kind of patients. A possible explanation is that these models fail

in understanding the complex relationships among variables in rheumatic patients. Moreover, they were developed many years ago and treatments have improved since then, so they probably are not well calibrated anymore. Including more recent populations and more novel biomarkers are two possible ways to face this issue. In this study we have evaluated the potential of ML approach to predict CV risk in rheumatic patients' cohorts. This an innovative approach, in fact, now, no similar applications exist. An automatic system based on Python programming language was developed, that improved traditional risk scores performances and that can be easily extended to other pathologies as well as to different kind of inflammatory arthritis. Moreover, it contains an easy tool to understand the importance of variables (potentially inside every type of pathologies). Looking at the results, we can see that ML approach demonstrated at least comparable results with traditional risk predictors, with the improvement in the sensitivity of PsA patient's classification, with respect to FRS. A big advantage in using ML methods is that they do not need to define cut-offs on predictions: they already produce a binary output. This can be useful in healthcare, because it simplifies the procedures. However, a cut-off may be imposed by unbalancing the training dataset during the learning phase of a model. We provide also a look into the variables weights of ML, by means of RF's importances, trying to understand results deeply and not only presenting them in the typical black box form. Variables which present higher weights should be included in further development of the algorithms.

The biggest limitation of this study is represented by the dataset's dimensions. 100-200 patients for each rheumatic group (PsA, AS and SLE) with only about 15% of CV events in each group is too small to allow the training and validation of a ML algorithm. Basing on this preliminary study we suggest that a dataset of about 500 or 1k patients (15% CV events) might be enough to allow training and validation of solid ML algorithms specific for the considered pathologies.

This work opens the way to personalized medicine and patient centered medicine, allowing development of models which are specific to group of patients and can assist doctors in the diagnostic and therapeutic process. A future work development can be the implementation of an online platform in which specialists can consult different score results.

5.6 Supporting Information

Table S10 - Performance metrics for cut-off points in cardiovascular risk scores applied to PsA patients. True cases represent all cardiovascular events that occurred, using the criteria for each individual risk score. The number of positive tests is the number of patients who are classified as being at intermediate-high or high risk for cardiovascular disease. OR, odds ratio.

	True cases (n)	Positive tests (n)	True positive (n)	False positive (n)	False negative (n)	True negative (n)	Total	Sensitivity (%)	Specificity (%)	Accuracy (%)	OR (%)
FRS											
> 10%	18	27	6	21	12	107	155	33.3	83.6	77.4	2.5
> 20%	18	7	3	4	15	124	155	16.7	96.9	87.0	6.2
CUORE											
> 10%	15	28	10	18	5	122	155	66.7	87.1	85.2	13.6
> 20%	15	9	3	6	12	134	155	20.0	95.7	88.4	5.6
SCORE											
> 1%	17	63	13	50	4	88	155	76.5	63.8	65.2	5.7
> 5%	17	5	1	4	16	134	155	5.9	97.1	87.1	2.1
FRS*1.5											
> 10%	18	48	13	35	5	93	155	72.2	72.7	72.6	6.9
> 20%	18	16	5	11	13	117	155	27.8	91.4	83.6	4.1
CUORE*1.5											
> 10%	15	42	11	31	4	109	155	73.3	77.9	77.4	9.7
> 20%	15	17	6	11	9	129	155	40.0	92.1	87.1	7.8
SCORE*1.5											
> 1%	17	112	16	96	1	42	155	94.1	30.4	37.4	7.0
> 5%	17	20	6	14	11	124	155	35.3	89.9	83.9	4.8

Table S11 - Performance metrics for cut-off points in cardiovascular risk scores applied to general patients. True cases represent all cardiovascular events that occurred, using the criteria for each individual risk score. The number of positive tests is the number of patients who are classified as being at intermediate-high or high risk for cardiovascular disease. OR, odds ratio.

	True cases (n)	Positive tests (n)	True positive (n)	False positive (n)	False negative (n)	True negative (n)	Total	Sensitivity (%)	Specificity (%)	Accuracy (%)	OR (%)
FRS											
> 10%	557	1900	436	1464	121	1637	3658	78.3	52.8	56.7	4.0
> 20%	557	895	276	619	281	2482	3658	49.6	80.0	75.4	3.9
CUORE											
> 10%	557	413	155	258	402	2843	3658	27.8	91.7	82.0	4.2
> 20%	557	106	52	54	505	3047	3658	9.3	98.3	84.7	5.8
SCORE											
> 1%	557	1109	323	786	234	2315	3658	58.0	74.7	72.1	4.1
> 5%	557	195	79	116	478	2985	3658	14.2	96.3	83.8	4.3

Table S12 - Performance metrics for machine learning algorithms applied to PsA patients. True cases represent all cardiovascular events that occurred, using the criteria for each individual risk score. The number of positive tests is the number of patients who are classified as being at intermediate-high or high risk for cardiovascular disease. OR, odds ratio.

	True cases (n)	**Positive tests (n)**	**True positive (n)**	**False positive (n)**	**False negative (n)**	**True negative (n)**	**Total**	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**	**OR (%)**
SVM	21	61	18	43	3	91	155	85.7	67.9	70.3	12.7
RF	21	55	18	37	3	97	155	85.7	72.4	74.2	15.7
KNN	21	71	19	52	2	82	155	90.5	61.2	65.2	15.0

6. References

Agca, R. *et al.* (2017) 'EULAR recommendations for cardiovascular disease risk management in patients with rheumatoid arthritis and other forms of inflammatory joint disorders: 2015/2016 update', *Annals of the Rheumatic Diseases*, 76(1), pp. 17–28. doi: 10.1136/annrheumdis-2016-209775.

D'Agostino, M. A., Palazzi, C. and Olivieri, I. (no date) 'Entheseal involvement.', *Clinical and experimental rheumatology*, 27(4 Suppl 55), pp. S50-5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19822046.

Ambale-Venkatesh, B. *et al.* (2017) 'Cardiovascular Event Prediction by Machine Learning', *Circulation Research*, 121(9), pp. 1092–1101. doi: 10.1161/CIRCRESAHA.117.311312.

Angermueller, C. *et al.* (2016) 'Deep learning for computational biology', *Molecular Systems Biology*, 12(7), p. 878. doi: 10.15252/msb.20156651.

Artigao-Rodenas, L. M. *et al.* (2013) 'Framingham Risk Score for Prediction of Cardiovascular Diseases: A Population-Based Study from Southern Europe', *PLoS ONE*. Edited by Y. Song, 8(9), p. e73529. doi: 10.1371/journal.pone.0073529.

Arts, E. E. A. *et al.* (2015) 'Performance of four current risk algorithms in predicting cardiovascular events in patients with early rheumatoid arthritis', *Annals of the Rheumatic Diseases*, 74(4), pp. 668–674. doi: 10.1136/annrheumdis-2013-204024.

Banerjee, P. and Preissner, R. (2018) 'BitterSweetForest: A Random Forest Based Binary Classifier to Predict Bitterness and Sweetness of Chemical Compounds', *Frontiers in Chemistry*, 6. doi: 10.3389/fchem.2018.00093.

Bastuji-Garin, S. *et al.* (2002) 'The Framingham prediction rule is not valid in a European population of treated hypertensive patients.', *Journal of hypertension*, 20(10), pp. 1973–80. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12359975.

Baxt, W. G. (1991) 'Use of an artificial neural network for the diagnosis of myocardial infarction.', *Annals of internal medicine*, 115(11), pp. 843–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1952470.

Bengtsson, K. *et al.* (2017) 'Are ankylosing spondylitis, psoriatic arthritis and undifferentiated

spondyloarthritis associated with an increased risk of cardiovascular events? A prospective nationwide population-based cohort study.', *Arthritis research & therapy*, 19(1), p. 102. doi: 10.1186/s13075-017-1315-z.

Benhamou, M., Gossec, L. and Dougados, M. (2010) 'Clinical relevance of C-reactive protein in ankylosing spondylitis and evaluation of the NSAIDs/coxibs' treatment effect on C-reactive protein', *Rheumatology*, 49(3), pp. 536–541. doi: 10.1093/rheumatology/kep393.

Brisimi, T. S. *et al.* (2018) 'Federated learning of predictive models from federated Electronic Health Records', *International Journal of Medical Informatics*, 112, pp. 59–67. doi: 10.1016/j.ijmedinf.2018.01.007.

Chen, X.-W. and Liu, M. (2005) 'Prediction of protein-protein interactions using random decision forest framework', *Bioinformatics*, 21(24), pp. 4394–4400. doi: 10.1093/bioinformatics/bti721.

Conroy, R. (2003) 'Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project', *European Heart Journal*, 24(11), pp. 987–1003. doi: 10.1016/S0195-668X(03)00114-3.

Cooksey, R. *et al.* (2018) 'Cardiovascular risk factors predicting cardiac events are different in patients with rheumatoid arthritis, psoriatic arthritis, and psoriasis', *Seminars in Arthritis and Rheumatism*, 48(3), pp. 367–373. doi: 10.1016/j.semarthrit.2018.03.005.

Cox, D. R. (1972) 'Regression Models and Life-Tables', *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), pp. 187–202. doi: 10.1111/j.2517-6161.1972.tb00899.x.

Crowson, C. S. *et al.* (2012) 'Usefulness of Risk Scores to Estimate the Risk of Cardiovascular Disease in Patients With Rheumatoid Arthritis', *The American Journal of Cardiology*, 110(3), pp. 420–424. doi: 10.1016/j.amjcard.2012.03.044.

Cui, S. *et al.* (2017) 'Plasma Phospholipids and Sphingolipids Identify Stent Restenosis After Percutaneous Coronary Intervention', *JACC: Cardiovascular Interventions*, 10(13), pp. 1307–1316. doi: 10.1016/j.jcin.2017.04.007.

D'Agostino, R. B. *et al.* (2008) 'General cardiovascular risk profile for use in primary care: the Framingham Heart Study.', *Circulation*, 117(6), pp. 743–53. doi: 10.1161/CIRCULATIONAHA.107.699579.

Damen, J. A. A. G. *et al.* (2016) 'Prediction models for cardiovascular disease risk in the general population: systematic review', *BMJ*, p. i2416. doi: 10.1136/bmj.i2416.

Detrano, R. *et al.* (1989) 'International application of a new probability algorithm for the diagnosis of coronary artery disease.', *The American journal of cardiology*, 64(5), pp. 304–10. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2756873.

Doria, A. *et al.* (2006) 'Long-Term Prognosis and Causes of Death in Systemic Lupus Erythematosus', *The American Journal of Medicine*, 119(8), pp. 700–706. doi: 10.1016/j.amjmed.2005.11.034.

Doukaki, S., Caputo, V. and Bongiorno, M. R. (2013) 'Psoriasis and Cardiovascular Risk: Assessment by CUORE Project Risk Score in Italian Patients.', *Dermatology research and practice*, 2013, p. 389031. doi: 10.1155/2013/389031.

Efron, B. and Tibshirani, R. (1997) 'Improvements on Cross-Validation: The 632+ Bootstrap Method', *Journal of the American Statistical Association*, 92(438), pp. 548–560. doi: 10.1080/01621459.1997.10474007.

Ernste, F. C. *et al.* (2015) 'Cardiovascular Risk Profile at the Onset of Psoriatic Arthritis: A Population-Based Cohort Study', *Arthritis Care & Research*, 67(7), pp. 1015–1021. doi: 10.1002/acr.22536.

Esteva, A. *et al.* (2017) 'Dermatologist-level classification of skin cancer with deep neural networks', *Nature*, 542(7639), pp. 115–118. doi: 10.1038/nature21056.

De Fauw, J. *et al.* (2018) 'Clinically applicable deep learning for diagnosis and referral in retinal disease', *Nature Medicine*, 24(9), pp. 1342–1350. doi: 10.1038/s41591-018-0107-6.

Ferri, F. J. *et al.* (1994) 'Comparative study of techniques for large-scale feature selection* *This work was suported by a SERC grant GR/E 97549. The first author was also supported by a FPI grant from the Spanish MEC, PF92 73546684', in, pp. 403–413. doi: 10.1016/B978-0-444-81892-8.50040-7.

Freidman, J. H., Bentley, J. L. and Finkel, R. A. (1977) 'An Algorithm for Finding Best Matches in Logarithmic Expected Time', *ACM Transactions on Mathematical Software*, 3(3), pp. 209–226. doi: 10.1145/355744.355745.

Galton, F. (1886) ‘Regression Towards Mediocrity in Hereditary Stature.’, *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15, p. 246. doi: 10.2307/2841583.

Gladman, D. *et al.* (1996) ‘The development and initial validation of the Systemic Lupus International Collaborating Clinics/American College of Rheumatology damage index for systemic lupus erythematosus.’, *Arthritis and rheumatism*, 39(3), pp. 363–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8607884.

Goldstein, B. A., Navar, A. M. and Carter, R. E. (2016) ‘Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges’, *European Heart Journal*, p. ehw302. doi: 10.1093/eurheartj/ehw302.

Gudadhe, M., Wankhade, K. and Dongre, S. (2010) ‘Decision support system for heart disease based on support vector machine and Artificial Neural Network’, in *2010 International Conference on Computer and Communication Technology (ICCCT)*. IEEE, pp. 741–745. doi: 10.1109/ICCCT.2010.5640377.

Gunaydin, Z. Y. *et al.* (2015) ‘Comparison of the Framingham risk and score models in predicting the presence and severity of coronary artery disease considering SYNTAX score’, *The Anatolian Journal of Cardiology*. doi: 10.5152/AnatolJCardiol.2015.6317.

Hajifathalian, K. *et al.* (2015) ‘A novel risk score to predict cardiovascular disease risk in national populations (Globorisk): a pooled analysis of prospective cohorts and health examination surveys’, *The Lancet Diabetes & Endocrinology*, 3(5), pp. 339–355. doi: 10.1016/S2213-8587(15)00081-9.

Hopkinson, N. D., Doherty, M. and Powell, R. J. (1994) ‘Clinical features and race-specific incidence/prevalence rates of systemic lupus erythematosus in a geographically complete cohort of patients.’, *Annals of the Rheumatic Diseases*, 53(10), pp. 675–680. doi: 10.1136/ard.53.10.675.

Hu, B. *et al.* (2018) ‘Machine learning algorithms based on signals from a single wearable inertial sensor can detect surface- and age-related differences in walking’, *Journal of Biomechanics*, 71, pp. 37–42. doi: 10.1016/j.jbiomech.2018.01.005.

Jackson, R. *et al.* (2005) ‘Treatment with drugs to lower blood pressure and blood cholesterol based on an individual’s absolute cardiovascular risk’, *The Lancet*, 365(9457), pp. 434–441. doi: 10.1016/S0140-6736(05)17833-7.

Johnson, K. W. *et al.* (2018) 'Artificial Intelligence in Cardiology', *Journal of the American College of Cardiology*, 71(23), pp. 2668–2679. doi: 10.1016/j.jacc.2018.03.521.

Kaharamanli, H. and Allahverdi, N. (2008) 'Design of a hybrid system for the diabetes and heart diseases', *Expert Systems with Applications*, 35(1–2), pp. 82–89. doi: 10.1016/j.eswa.2007.06.004.

Kim, J., Kang, U. and Lee, Y. (2017) 'Statistics and Deep Belief Network-Based Cardiovascular Risk Prediction', *Healthcare Informatics Research*, 23(3), p. 169. doi: 10.4258/hir.2017.23.3.169.

Kinsella, T. D., Johnson, L. G. and Ian, R. (1974) 'Cardiovascular manifestations of ankylosing spondylitis.', *Canadian Medical Association journal*, 111(12), pp. 1309–11. Available at: http://www.ncbi.nlm.nih.gov/pubmed/4442013.

LeCun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *Nature*, 521(7553), pp. 436–444. doi: 10.1038/nature14539.

Li, Q., Rajagopalan, C. and Clifford, G. D. (2014) 'A machine learning approach to multi-level ECG signal quality classification', *Computer Methods and Programs in Biomedicine*, 117(3), pp. 435–447. doi: 10.1016/j.cmpb.2014.09.002.

McCulloch, W. S. and Pitts, W. (1990) 'A logical calculus of the ideas immanent in nervous activity', *Bulletin of Mathematical Biology*, 52(1–2), pp. 99–115. doi: 10.1007/BF02459570.

McMahon, M. and Hahn, B. H. (2007) 'Atherosclerosis and systemic lupus erythematosus — mechanistic basis of the association', *Current Opinion in Immunology*, 19(6), pp. 633–639. doi: 10.1016/j.coi.2007.11.001.

Mease, P. J. (2005) 'Psoriatic arthritis assessment tools in clinical trials', *Annals of the Rheumatic Diseases*, 64(suppl_2), pp. ii49–ii54. doi: 10.1136/ard.2004.034165.

Miller, D. D. and Brown, E. W. (2018) 'Artificial Intelligence in Medical Practice: The Question to the Answer?', *The American Journal of Medicine*, 131(2), pp. 129–133. doi: 10.1016/j.amjmed.2017.10.035.

Narain, R., Saxena, S. and Goyal, A. (2016) 'Cardiovascular risk prediction: a comparative study of Framingham and quantum neural network based approach', *Patient Preference and Adherence*, Volume 10, pp. 1259–1270. doi: 10.2147/PPA.S108203.

Narula, S. *et al.* (2016) 'Machine-Learning Algorithms to Automate Morphological and Functional Assessments in 2D Echocardiography', *Journal of the American College of Cardiology*, 68(21), pp. 2287–2295. doi: 10.1016/j.jacc.2016.08.062.

Navarini, L. *et al.* (2018) 'Performances of five risk algorithms in predicting cardiovascular events in patients with Psoriatic Arthritis: An Italian bicentric study', *PLOS ONE*. Edited by C. Zito, 13(10), p. e0205506. doi: 10.1371/journal.pone.0205506.

Obermeyer, Z. and Emanuel, E. J. (2016) 'Predicting the Future — Big Data, Machine Learning, and Clinical Medicine', *New England Journal of Medicine*, 375(13), pp. 1216–1219. doi: 10.1056/NEJMp1606181.

Palaniappan, S. and Awang, R. (2008) 'Intelligent heart disease prediction system using data mining techniques', in *2008 IEEE/ACS International Conference on Computer Systems and Applications*. IEEE, pp. 108–115. doi: 10.1109/AICCSA.2008.4493524.

Panch, T., Szolovits, P. and Atun, R. (2018) 'Artificial intelligence, machine learning and health systems', *Journal of Global Health*, 8(2). doi: 10.7189/jogh.08.020303.

Park, C., Took, C. C. and Seong, J.-K. (2018) 'Machine learning in biomedical engineering', *Biomedical Engineering Letters*, 8(1), pp. 1–3. doi: 10.1007/s13534-018-0058-3.

Pham, T. *et al.* (2017) 'Predicting healthcare trajectories from medical records: A deep learning approach', *Journal of Biomedical Informatics*, 69, pp. 218–229. doi: 10.1016/j.jbi.2017.04.001.

Pouriyeh, S. *et al.* (2017) 'A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease', in *2017 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, pp. 204–207. doi: 10.1109/ISCC.2017.8024530.

Ritchlin, C. T., Colbert, R. A. and Gladman, D. D. (2017) 'Psoriatic Arthritis', *New England Journal of Medicine*. Edited by D. L. Longo, 376(10), pp. 957–970. doi: 10.1056/NEJMra1505557.

Roifman, I. *et al.* (2011) 'Chronic Inflammatory Diseases and Cardiovascular Risk: A Systematic Review', *Canadian Journal of Cardiology*, 27(2), pp. 174–182. doi: 10.1016/j.cjca.2010.12.040.

Rosenblatt, F. (1958) 'The perceptron: A probabilistic model for information storage and organization in the brain.', *Psychological Review*, 65(6), pp. 386–408. doi: 10.1037/h0042519.

Ross, R. (1999) ‘Atherosclerosis — An Inflammatory Disease’, *New England Journal of Medicine*. Edited by F. H. Epstein, 340(2), pp. 115–126. doi: 10.1056/NEJM199901143400207.

Salvarani, C. *et al.* (1995) ‘Prevalence of psoriatic arthritis in Italian psoriatic patients.’, *The Journal of rheumatology*, 22(8), pp. 1499–503. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7473473.

Salvarani, C., Gabriel, S. and Hunder, G. G. (1996) ‘Distal extremity swelling with pitting edema in polymyalgia rheumatica. Report of nineteen cases’, *Arthritis & Rheumatism*, 39(1), pp. 73–80. doi: 10.1002/art.1780390110.

Samuel, A. L. (1959) ‘Some Studies in Machine Learning Using the Game of Checkers’, *IBM Journal of Research and Development*, 3(3), pp. 210–229. doi: 10.1147/rd.33.0210.

Sannino, G. and De Pietro, G. (2011) ‘A smart context-aware mobile monitoring system for heart patients’, in *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*. IEEE, pp. 655–695. doi: 10.1109/BIBMW.2011.6112448.

Schnyer, D. M. *et al.* (2017) ‘Evaluating the diagnostic utility of applying a machine learning algorithm to diffusion tensor MRI measures in individuals with major depressive disorder’, *Psychiatry Research: Neuroimaging*, 264, pp. 1–9. doi: 10.1016/j.pscychresns.2017.03.003.

Shah, S. J. *et al.* (2015) ‘Phenomapping for Novel Classification of Heart Failure With Preserved Ejection Fraction’, *Circulation*, 131(3), pp. 269–279. doi: 10.1161/CIRCULATIONAHA.114.010637.

Shoenfeld, Y., Sherer, Y. and Harats, D. (2001) ‘Artherosclerosis as an infectious, inflammatory and autoimmune disease.’, *Trends in immunology*, 22(6), pp. 293–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11419409.

Sutton, E. J., Davidson, J. E. and Bruce, I. N. (2013) ‘The Systemic Lupus International Collaborating Clinics (SLICC) damage index: A systematic literature review’, *Seminars in Arthritis and Rheumatism*, 43(3), pp. 352–361. doi: 10.1016/j.semarthrit.2013.05.003.

Tajik, A. J. (2016) ‘Machine Learning for Echocardiographic Imaging’, *Journal of the American College of Cardiology*, 68(21), pp. 2296–2298. doi: 10.1016/j.jacc.2016.09.915.

Tsokos, G. C. (2011) ‘Systemic Lupus Erythematosus’, *New England Journal of Medicine*,

365(22), pp. 2110–2121. doi: 10.1056/NEJMra1100359.

Tu, J. V. and Guerriere, M. R. J. (1993) 'Use of a Neural Network as a Predictive Instrument for Length of Stay in the Intensive Care Unit Following Cardiac Surgery', *Computers and Biomedical Research*, 26(3), pp. 220–229. doi: 10.1006/cbmr.1993.1015.

Wang, R. and Ward, M. M. (2018) 'Epidemiology of axial spondyloarthritis', *Current Opinion in Rheumatology*, 30(2), pp. 137–143. doi: 10.1097/BOR.0000000000000475.

Weng, S. F. *et al.* (2017) 'Can machine-learning improve cardiovascular risk prediction using routine clinical data?', *PLOS ONE*. Edited by B. Liu, 12(4), p. e0174944. doi: 10.1371/journal.pone.0174944.

Willems, J. L. *et al.* (1991) 'The Diagnostic Performance of Computer Programs for the Interpretation of Electrocardiograms', *New England Journal of Medicine*, 325(25), pp. 1767–1773. doi: 10.1056/NEJM199112193252503.

Wolpert, D. H. (1996) 'The Lack of A Priori Distinctions Between Learning Algorithms', *Neural Computation*, 8(7), pp. 1341–1390. doi: 10.1162/neco.1996.8.7.1341.

Wolpert, D. H. and Macready, W. G. (1997) 'No free lunch theorems for optimization', *IEEE Transactions on Evolutionary Computation*, 1(1), pp. 67–82. doi: 10.1109/4235.585893.

Yaniv, G. *et al.* (2015) 'A volcanic explosion of autoantibodies in systemic lupus erythematosus: A diversity of 180 different antibodies found in SLE patients', *Autoimmunity Reviews*, 14(1), pp. 75–79. doi: 10.1016/j.autrev.2014.10.003.

Yim, K. M. and Armstrong, A. W. (2017) 'Updates on cardiovascular comorbidities associated with psoriatic diseases: epidemiology and mechanisms', *Rheumatology International*, 37(1), pp. 97–105. doi: 10.1007/s00296-016-3487-2.