

POLITECNICO DI MILANO

School of Industrial and Information Engineering

Master of Science in Biomedical Engineering



**EXPLAINABLE ARTIFICIAL INTELLIGENCE:  
REVIEW AND APPLICATIONS IN MEDICAL  
FIELD**

Supervisor:      Pietro Cerveri

Thesis of  
Simone Fiorani  
Student ID: 914800

Academic Year 2020-2021



# Abstract

Technological progress has led in the last decade to the development of increasingly intelligent and autonomous artificial intelligence systems, but at the same time also complex and difficult to understand. In some areas, the opacity of these systems calls into question their application, despite the results provided by these technologies are undoubtedly useful: this reluctance is particularly high in the medical sector, in which stringent requirements limit the use of AI technologies both from the ethical and legal point of view and from the application one.

This thesis aims to present one of the possible solutions to the problem of opacity of AI technologies, called Explainable Artificial Intelligence. These tools aim to be an evolution of the classic AI techniques but increasing their level of transparency: in doing so these instruments become more comprehensible, so more controllable, correctable, and reliable. Following this path, you can implement these technologies even in the most reluctant sectors, such as the medical, because their work is more understandable to both developers, and end users.

This review allows to say that, despite the development of Explainable Artificial Intelligence is still in an early stage, the potential is very high: the benefits that can be derived from the explanations and justifications of the work done by automated systems allow to enrich the knowledge that one has both on the development of these tools, and on the areas in which they are applied. There are of course problems and limitations, but with a joint effort of developers, industry experts and end-users, you can bypass these problems and increase the integration of AI in areas still reluctant.

# Abstract

Il progresso tecnologico ha portato nell'ultimo decennio allo sviluppo di sistemi di intelligenza artificiale sempre più intelligenti e autonomi, ma allo stesso tempo anche complessi e di difficile comprensione. In alcune aree, l'opacità di questi sistemi mette in discussione la loro applicazione, nonostante i risultati forniti da queste tecnologie siano indubbiamente utili: questa avversione è particolarmente elevata nel settore medico, in cui requisiti rigorosi limitano l'uso delle tecnologie dell'AI sia dal punto di vista etico e giuridico, che da quello applicativo.

Questa tesi mira a presentare una delle possibili soluzioni al problema dell'opacità delle tecnologie AI, chiamata Intelligenza Artificiale Spiegabile. Questi strumenti mirano ad essere un'evoluzione delle classiche tecniche di AI ma aumentandone il livello di trasparenza: così facendo questi strumenti diventano più comprensibili, quindi più controllabili, correggibili e affidabili. Seguendo questo percorso, è possibile implementare queste tecnologie anche nei settori più riluttanti, come quello medico, perché il loro lavoro è più comprensibile sia per gli sviluppatori, sia per gli utenti finali.

Questo lavoro permette di dire che, nonostante lo sviluppo dell'Intelligenza Artificiale Spiegabile sia ancora in una fase iniziale, il suo potenziale è molto alto: i benefici che possono derivare dalle spiegazioni e giustificazioni del lavoro svolto dai sistemi automatizzati permettono di arricchire le conoscenze che si hanno sia sullo sviluppo di questi strumenti, sia sulle aree in cui vengono applicati. Naturalmente ci sono problemi e limitazioni, ma con uno sforzo congiunto di sviluppatori, esperti del settore e utenti finali, è possibile aggirare questi problemi e aumentare l'integrazione di AI in aree ancora restie.



# Sommario

## Introduzione

Nella cultura popolare, il termine "intelligenza artificiale" è di solito usato in riferimento ad una tecnologia futuristica, lontana dal mondo in cui viviamo, come una strana entità superiore proveniente da un film di fantascienza. Al contrario, nell'ultimo decennio diversi tipi di AI sono utilizzati nei settori industriali e di ricerca, ma anche nella nostra vita quotidiana: auto a guida autonoma, assistenti domestici per gestire elettrodomestici, assistenti negli smartphone, e così via.

La diffusione di questi strumenti è così elevata che anche le istituzioni hanno iniziato a regolamentare l'uso di questi sistemi. La Commissione Europea, ad esempio, negli ultimi anni ha lanciato una massiccia campagna per regolamentare l'uso di sistemi automatizzati, soprattutto dal punto di vista del rispetto della privacy e dell'etica umana nell'uso di questi strumenti: documenti come il "Regolamento Generale sulla Protezione dei Dati" o "Ethics guidelines for trustworthy AI" devono essere viste in questa luce.

In medicina e nei sistemi sanitari, l'AI è utilizzata per assistere gli esperti nel prendere decisioni o gestire queste istituzioni: algoritmi di deep learning sono utilizzati per analizzare le immagini mediche e assistere i medici nella diagnosi, modelli di machine learning per aiutare l'amministrazione a capire come è possibile ottimizzare i costi e migliorare le strutture e i servizi forniti dal sistema sanitario. Le potenziali applicazioni dei sistemi automatizzati nell'ambiente medico, possono portare ad un rinnovamento del settore, in particolare aumentando la velocità dell'analisi dei dati e l'assistenza ai processi decisionali, mantenendo un alto livello di accuratezza.

Tuttavia, l'enorme potenziale dell'IA è anche accompagnato da diverse problematiche, in particolare per quanto riguarda la trasparenza di questi strumenti nel loro lavoro: infatti, come vedremo più avanti, la maggior parte di queste tecnologie operano in un modo non del tutto comprensibile, sia perché le aziende tendono a mantenere segrete le loro invenzioni, sia perché il design di queste tecnologie ha ancora dei lati oscuri. Questi problemi sono particolarmente limitanti

nei settori in cui la trasparenza del proprio lavoro è fondamentale: questa categoria comprende il principale settore di interesse di questo lavoro, quello medico. Nei sistemi sanitari, infatti, la diffusione delle tecnologie automatizzate deve necessariamente scontrarsi con ostacoli legali, perché tutto ciò che riguarda il settore medico deve rispettare regole severe, con l'obiettivo di aumentare al massimo la sicurezza di qualsiasi strumentazione.

## **Obiettivi**

Lo scopo di questa tesi è proprio quello di analizzare una delle possibili tecniche il cui obiettivo è quello di appianare i limiti offerti dall'attuale stato dell'arte dell'AI, vale a dire l'AI spiegabile (Explainable AI, XAI), e mostrare come questa tecnologia può essere sfruttata nel settore medico, che forse più di ogni altro ha bisogno di uno strumento come questo. Le pagine seguenti offriranno un percorso introduttivo a questa evoluzione tecnologica: una presentazione iniziale e generale delle principali tecniche dell'IA classica, che si basano principalmente su due approcci, il Machine Learning e il Deep Learning. A questo capitolo segue la presentazione della versione spiegabile delle AI: come è possibile implementare le spiegazioni, quali di queste devono essere scelte da uno sviluppatore a seconda del suo campo di applicazione, qual è lo stato attuale dell'arte delle XAI. Infine, parleremo di come le XAI sono implementate nel sistema sanitario, con alcuni esempi pratici di opere che sfruttano questa tecnologia: saranno mostrate quattro diverse opere, in cui l'AI spiegabile viene sfruttata in campo medico, con applicazioni anche molto attuali come quelle sul Covid-19 o sugli anticorpi monoclonali, con questi ultimi che hanno ricevuto una grande spinta proprio dallo stato pandemico attualmente presente.

## **Conclusioni**

Le XAI offrono sicuramente una validissima alternativa alle classiche tecniche di intelligenza artificiale, in particolare in quegli ambiti come quello medico nel quale i requisiti chiesti alle strumentazioni sono estremamente esigenti. La possibilità di ottenere spiegazioni e giustificazione delle scelte fatte dai sistemi automatizzati

permette di aumentare la comprensibilità di questi strumenti, e allo stesso tempo migliorare la confidenza con cui il personale medico li utilizza, offre la possibilità di aumentare il bagaglio culturale sia dal punto di vista dello sviluppo delle AI, che dal punto di vista della conoscenza generale del fenomeno sotto analisi. Queste possibilità sono accompagnate da alcune problematiche, e sarà compito della comunità scientifica fare uno sforzo comune per migliorare questa tecnologia che si presenta come uno dei principali candidati a dare il giusto trampolino di lancio alla definitiva diffusione delle tecniche AI anche in quei settori più restii ad accettarle. A questo sforzo si dovrà accompagnare anche una maggiore elasticità da parte del settore medico, in particolare aggiornando il personale già al lavoro e progettando un percorso educativo delle future generazioni che integri maggiormente queste tecnologie, per poi poterle maneggiare al meglio nella pratica quotidiana.





# Contents

<b>1. Introduction.....</b>	<b>1</b>
<b>2. Artificial Intelligence .....</b>	<b>3</b>
<b>2.1 Machine Learning.....</b>	<b>4</b>
2.1.1 Supervised Learning.....	8
2.1.2 Unsupervised Learning .....	13
<b>2.2 Deep Learning .....</b>	<b>17</b>
2.2.1 From Artificial Neuron to Artificial Neural Network.....	18
2.2.2 Training an ANN.....	30
2.2.3 The convolutional neural network.....	37
<b>2.3 AI in medicine and healthcare .....</b>	<b>44</b>
<b>3. Explainable Artificial Intelligence .....</b>	<b>48</b>
<b>3.1 Why use an XAI .....</b>	<b>54</b>
<b>3.2 How to implement explainability.....</b>	<b>58</b>
3.2.1 Explainability depending on complexity.....	59
3.2.2 Explainability depending on scope .....	60
3.2.3 Explainability depending on model dependency .....	62
<b>4. XAI in medicine.....</b>	<b>65</b>
<b>4.1 XAI application .....</b>	<b>68</b>
4.1.1 COVID-19.....	68
4.1.2 Monoclonal antibodies .....	73
4.1.3 Heart failure.....	77
4.1.4 Breast cancer diagnosis .....	82
<b>4.2 Limitations and challenges.....</b>	<b>86</b>
<b>5. Conclusions.....</b>	<b>91</b>
<b>Bibliography.....</b>	<b>94</b>



# List of figures

FIGURE 1: AI HIERARCHY .....	5
FIGURE 2: STRUCTURE OF SUPERVISED AND UNSUPERVISED LEARNING.....	8
FIGURE 3: EXAMPLE OF THE STARTING POINT OF A SVM LEARNING METHODS .....	11
FIGURE 4: EXAMPLE OF SEPARATING HYPERPLANE IN THE SVM PROCESS.....	12
FIGURE 5: EXAMPLES OF NONLINEAR SEPARATION IN SVM .....	13
FIGURE 6: EXAMPLE OF APPLICATION OF THE K-MEANS ALGORITHM .....	15
FIGURE 7: EXAMPLE OF A HUMAN NEURON.....	19
FIGURE 8: EXAMPLE OF NEURON PRINCIPLE.....	20
FIGURE 9: EXAMPLES OF ACTIVATION FUNCTIONS .....	21
FIGURE 10: STRUCTURE OF AN ARTIFICIAL NEURON THAT CAN DECIDE TO GO FOR A WALK OR NOT .....	23
FIGURE 11: I/O RELATIONSHIP OF AN AND PORT AND VISUALIZATION OF THE OUTPUT IN THE INPUTS SPACE.....	24
FIGURE 12: STRUCTURE OF AN ARTIFICIAL NEURON AS AN AND PORT.....	25
FIGURE 13: SEPARATION LINE BETWEEN POINTS .....	26
FIGURE 14: I/O RELATIONSHIP OF AN AND PORT AND VISUALIZATION OF THE OUTPUT IN THE INPUTS SPACE.....	26
FIGURE 15: TOPOLOGY OF THE NETWORK THAT PERFORM THE XOR OPERATION .....	27
FIGURE 16: EXAMPLE OF A DEEP NEURAL NETWORK .....	29
FIGURE 17: EXAMPLE OF PERCEPTRON WITH A HIDDEN LAYER .....	33
FIGURE 18: EXAMPLE OF CNN.....	38
FIGURE 19: EXAMPLE OF A FEATURE MAP.....	39
FIGURE 20: EXAMPLE OF A KERNEL MOVING ON AN IMAGE.....	40
FIGURE 21: EXAMPLE OF DIFFERENT FEATURE MAPS .....	41
FIGURE 22: EXAMPLE OF RELU FUNCTION.....	42
FIGURE 23: EXAMPLE OF APPLICATION OF MAX POOLING .....	43
FIGURE 24: EXAMPLE OF WHAT THE DEEPER LAYERS OF A CNN LEARN .....	44
FIGURE 25: DOUBLE DESCENT PHENOMENON .....	50
FIGURE 26: COMPARISON BETWEEN AI AND XAI .....	52
FIGURE 27: GOOGLE TREND FOR "EXPLAINABLE ARTIFICIAL INTELLIGENCE" OF THE LAST 5 YEARS .....	53

FIGURE 28: MAIN OBJECTIVES FOR AN XAI.....	54
FIGURE 29: REASONS TO USE AN XAI .....	55
FIGURE 30: SALIENCY MAPS IN CNN IMAGE ANALYSIS .....	61
FIGURE 31: EXAMPLE OF HEAT MAP .....	71
FIGURE 32: EXAMPLE OF LOW AND HIGH LEVEL OF EXPLANATIONS.....	72
FIGURE 33: FALSE NEGATIVE PATIENTS .....	73
FIGURE 34: KNOWLEDGE TRANSFER PROCESS .....	76
FIGURE 35: SCHEME OF THE INTERPRETABILITY APPROACH.....	76
FIGURE 36: EXAMPLE OF GAN.....	79
FIGURE 37: EXAMPLE OF AN AUTOENCODER NETWORK.....	79
FIGURE 38: EXAMPLE OF GVR.....	81
FIGURE 39: EXAMPLE OF LP FEATURE VISUALIZATION .....	83
FIGURE 40: EXAMPLE OF RADVIZ VISUALIZATION.....	84
FIGURE 41: RESULTS OF THE CLASSIFIERS.....	85
FIGURE 42: BASELINE CART STRUCTURE .....	85

**List of tables**

TABLE 1: EXAMPLE OF TABLE FOR ML..... 6

TABLE 2: EXAMPLE OF A TABLE FOR SUPERVISED LEARNING ..... 9

TABLE 3: EXAMPLE OF WHICH COULD BE THE DETERMINING ASPECTS THAT  
AFFECT THE FINAL DECISION ..... 23

# List of abbreviations

AI	Artificial Intelligence
NN	Neural network
XAI	Explainable Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
PR	Pattern Recognition
SL	Supervised Learning
UL	Unsupervised Learning
SVM	Support Vector Machine
ANN	Artificial Neural Network
FFNN	Feed-Forward Neural Network
CNN	Convolutional Neural Network
GDPR	General Data Protection Regulation
CDSS	Clinical Decision Support Systems
FDA	Food and Drugs Administration
AIFA	Agenzia Italiana del Farmaco
GAN	Generative Adversarial Networks





# ***1. Introduction***

In popular culture, the term “artificial intelligence” is usually used referring to futuristic technology, far away from the world where we live, like a strange superior entity from a sci-fi film. Conversely, in the last decade different types of AI are used in industrial and research sectors, but also in our daily life: self-driving cars, smart devices to manage home appliances, assistants in smartphones, and so on.

The dissemination of these tools is so high that the institutions have also begun to regulate the use of these systems. The European Commission, for example, has in recent years launched a massive campaign to regulate the use of automated systems, especially from the point of view of respect for privacy and human ethics in the use of these tools: documents such as the “General Data Protection Regulation” or the “Ethics guidelines for trustworthy AI” [1] must be seen in this light.

A daily life example of this kind of technology is the home assistant that can be found in our houses: this AI system can control smart devices inside our houses like televisions and speakers, but also lighting systems or hobs. The level of smartness is so high, that interaction between tenants and device, allows the latter to learn the habits of inhabitants, and regulate for example a thermostat in complete autonomy without people’s control.

In the last decades, also governments are starting to apply AI in their work of managing cities or countries: face recognition made by neural networks (from now on NN) for security, machine learning applied in traffic control or in jurisprudence to help judges to give a right sentence.

In medicine and healthcare, AI are used to assist experts in take decisions or run a healthcare system: deep learning algorithms are used to analyse medical images to assist doctors in diagnosis, while machine learning methods to help the management to understand how is possible to optimize costs and improve the structures and services provided by the healthcare system. The potential applications of automated systems in the medical environment, can lead to a renewal of the sector, in particular

by increasing the speed of data analysis and assistance to decision-making processes, maintaining a high level of accuracy.

However, the enormous potential of AI is also accompanied by several issues, particularly in the area of the transparency of these instruments in their work: in fact, as we shall see below, most of these technologies operate in a way that is not entirely understandable, both because companies tend to keep their inventions secret, and because the design of these technologies still has dark sides. These problems are particularly limiting in areas where the transparency of one's own work is fundamental: this category includes the main area of interest of this work, the medical one. In health systems, in fact, the dissemination of automated technologies must necessarily face legal obstacles, because everything relating to the medical sector must comply with strict rules, with the aim of increasing the safety of any instrumentation to the maximum possible.

The aim of this thesis is precisely to analyse one of the possible techniques whose objective is to smooth the limitations offered by the current state of the art of AI, namely the Explainable AI (XAI), and show how this technology can be exploited in the medical sector, which perhaps more than any other needs a tool like this. The following pages will offer an introductory path to this technological evolution: an initial and general presentation of the main techniques of the classic AI, which are mainly based on two approaches: the Machine Learning and the Deep Learning method. To this chapter follows the presentation of the XAI version: how it is possible to implement explanations, which of these explanations must be chosen by a developer depending on his application field, which is the actual state of the art of XAI. Finally, we will discuss how the XAI are implemented in the healthcare system, with some practical examples of works that exploit this technology: four different works will be shown, in which the explainable AI are exploited in the medical field.

## 2. *Artificial Intelligence*

During the last decades, due to the constant increase of interest around the artificial intelligence, companies and manufacturers started to indicate in the name or in the description of their product the term AI, just to increase their attractiveness: this phenomenon is called *AI washing*, the idea of adding the label of AI to all and every software platform [2].

In reality, it is really complicated to give a unique definition of Artificial Intelligence, due to the huge number of algorithms, structures, and ways to develop one. Salehi and Burgueño [3] say that AI “refers to a machine’s ability to mimic the cognitive functions of humans to perform tasks in a smart manner.” In general, the goal of AI is to create machines that can solve problems with the accuracy and speed of computers, but in a way of “thinking”, so with an *intelligence*, that is proper of human beings.

A more technical definition of AI is given by Haenlein and Kaplan [4], who describe an AI as “a system’s ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation”. This description uses some truly significant words:

- An AI is a **system**, so the integration of different entities. These entities can be hardware resources, software, or a combination of both, and forms an architecture.
- This system must **achieve a specific goal and tasks**, so the different parts of this structure collaborate to reach a defined target.
- To do this, the AI needs the **ability to interpret external data**, so it can analyse information from the world and exploit them to achieve the task.
- An AI uses the information to **learn**, so it can modify its structure and update itself using a **flexible adaptation**, always to achieve the goal.

So, unifying the two descriptions, an AI can be described as a complex system in which, starting from a learning process of information, it develops the ability to interpret this data and exploit them to achieve a specific task in a way similar to what a human being would do.

Always [4], identify three main AI methods that are most diffused in the last decade: **Machine Learning (ML)**, **Pattern Recognition (PR)** and **Deep Learning (DL)**.

While ML and DL are usually accepted as AI paradigms, for PR it is not uncommon to find some authors that give different definitions of PR: Pavlidis [5] says that “problems of pattern recognition were generally lumped under the term *Artificial Intelligence*, although a more appropriate name might be *Machine Intelligence*”. With the term Machine Intelligence, it is meant when “machines are programmed with some (but not all) aspects of human intelligence, including learning, problem solving and prioritization. With these (limited) abilities, a machine can tackle a complex set of problems” [6]. So, it is not a stand-alone method, but rather one of the many problems that AI can solve.

Another definition of PR is given by Chao [7], which says “Pattern recognition is a process that taking in raw data and making an action based on the category of the pattern”, calling *pattern* a “entity vaguely defined, that could be given a name”. So, again PR it is not an algorithm, but a process that can be done in different ways, also by exploiting ML and DL.

For these reasons, in this chapter will be presented the ML and DL, as the most diffused paradigm of AI, trying to understand which are the differences between these two approaches, and their potential and limitations.

## 2.1 Machine Learning

The improvement of technology and the advent of internet has profoundly changed the way each person relates to the world: faster communications between people, easier access to information and the spread of sensitive data in social media or public services. In the last 20 years the production of data has raised at an incredible speed, and its collection led to the storage of a quantity of information never reached before. Just as an example, the European Commission in [8] says that “The volume of data produced in the world is growing rapidly, from 33 zettabytes in 2018 to an expected

175 zettabytes<sup>1</sup> in 2025”. It is an incredible amount of data, if compared to the storage capacity available just 20 years ago, represented by the floppy disk (2-4 MB) or a CD (600-700 MB): just to make it clearer, to collect one zettabyte we would need around  $10^{18}$  CDs.

All this available data gave the necessary thrust to the development of architectures to collect all these information and to the creation of algorithms able to analyse this huge amount of information: Machine Learning was born for this purpose.

Indeed, ML is a branch of AI that tries to develop systems analyse a huge amount of data and manipulate them to extract some characteristic features or to group population of data into homogeneous clusters. Nowadays, ML is widespread and implemented in almost every field of technology, research, economy, and supply chains.

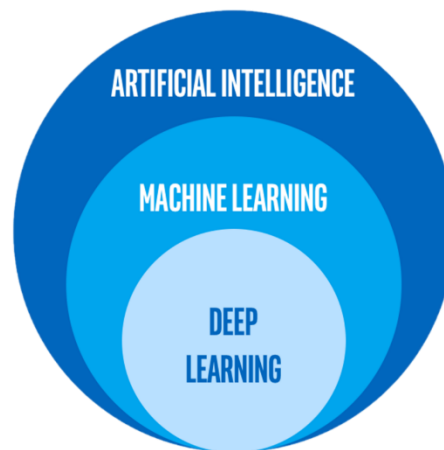


Figure 1: AI hierarchy

Cristoph Molnar [9] summarizes ML as a “set of methods that computers use to make and improve predictions or behaviours based on data”, while in [3] is said that “ML refers to the capability of computers to learn without being explicitly programmed”. Indeed, ML systems are based on their learning capability, exploiting different mathematical models that rule how to process the data: the only “programming” of

---

<sup>1</sup> 1 zettabyte is equal to  $10^{21}$  bytes, so about  $10^{12}$  gigabytes.

the system are the rules of the mathematical model, then the algorithm can learn which is the correlation between input and output in complete autonomy, so it is like the system is generating the program itself. From that point, it can continuously update itself, in a constant attempt to improve its work.

The word **data** has already been used a lot. Data is the basis from which all ML methods begin, and are summarised into **databases**, so collections of information usually correlated with the phenomenon under analysis. In ML, databases are in organized in the form of tables, and in order to **correctly train** an algorithm, it is better to have databases that are collections of information about the specific object of interest, or more generally of what we want to analyse.

In the case, for example, of medical statistics, we might be interested in which are the correlations between tumours and different aspects of human life: we would need a database that contains information on daily life of patient with tumours or not, like if they smoke or not, which is their diet, if they do sports or not.

Looking at the structure of Table 1, it is composed by three fundamental entities. On top of each column there is the **feature** or **attribute**, so a specific characteristic of the object of interest, for example if a patient is a smoker or how many hours of physical activity he/she practices daily. At the beginning of each row, there is the **record**, so the element of the table that contains the **values** associated to the feature, for example that particular patient called “Mario Rossi”; values could be **numerical** (numbers), **categorical** (words like the days in a week or gender, yes/no and so on) and in a table you may find only one type of value or a combination of both.

	Age	Smoker	Familiarity	Daily Sport Hrs
<i>Mario Rossi</i>	58	Yes	Yes	1
<i>Anna Bianchi</i>	49	No	No	3
<i>Piero Verdi</i>	73	No	Yes	1.5

Table 1: example of table for ML

So, if we would translate in phrases the first record of the table, we would say: “Mario Rossi is 58 years old, he is a smoker, he is a patient with familiarity, and he practises 1 hour of sport daily.

Companies invest a lot of effort and money to create these collections of data, using surveys, exploiting cookies online or just creating specific study-cases on the topic of interest. The reason is that data and ML (but more generally AI) allow to create models of almost every aspect of a phenomenon: so, more information about something is available, more accurate will be the analyses we do.

All these analyses start with the **learning process**, the most powerful ability of AI systems. Algorithms exploit databases as the “literature” that they “read” to learn something. There are different types of ML techniques, that differ for the goal they want to reach, the way to achieve the task and so on, but they can be at first divided in two main categories: the **supervised machine learning** and the **unsupervised machine learning**.

To understand the difference between these two major categories, it is necessary to take a step back. In Table 1 it had been shown an example of table, in which we had different features: we said that the **attributes** are particular aspects of the phenomenon under analysis, and one of these can be extracted and used as a **target** for the investigation. Using the same example of before, if we are studying the incidence rate of tumour in a population, we can consider a feature “tumour” significant for the study, because it gives a sort of “final result” to all the other features of the records: “*the patient has a tumour, because of these features have these values...*”, or “*x % of smoker has a tumor*”.

The presence or not of the target separates the learning process in *supervised* and *unsupervised*. In the first group are included the **classification** and **regression models**, while to the second belong the analysis with purpose **to identify regularities, similarities, and differences** in the data [10], like clustering or anomaly detection.

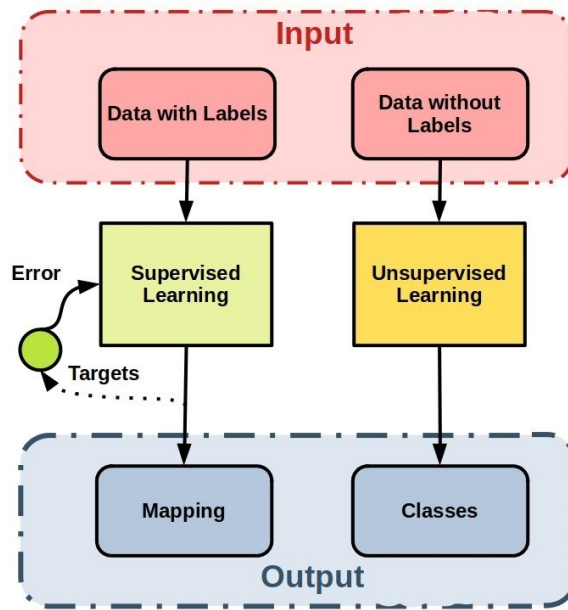


Figure 2: structure of Supervised and Unsupervised Learning.

As shown in Figure 2, the two learning procedures reach different tasks, and in a different way: the Supervised Learning (from now on SL) exploits *Data with Labels* (the label is the target attribute) as input, uses the *targets* to understand if he learnt a good model or not through an *error* parameter, and finally provides a *mapping* as output. Conversely, the Unsupervised Learning (UL from now on) uses *Data without Labels*, learns the model without an error parameter, and produces some *Classes* output.

In the next pages will be presented the two the methods, giving an example of a specific mathematical framework for both the approaches.

### 2.1.1 Supervised Learning

Under the term *Supervised Learning*, are grouped all the models that exploit the presence of the *target attribute* in their learning procedure. The goal is “learning a mapping between a set of input variables  $X$  and an output variable  $Y$  and applying this mapping to predict the outputs for unseen data” [11].



The input variables are the records from the database, accompanied with the values and the attributes: the table must also contain the output, which is the class label or target. For instance, an extension of Table 1, could be the addition of another attribute called “Tumour” as target of the analysis, which means “that the patient has cancer”. As shown in Table 2, the target could be numerical or categorical: the first one category leads to a regression problem, the latter into a classification one. In any case, the categorical values must be converted into numbers like 0/1 or an ordered sequence of numbers. For instance, the target YES/NO in the table must be converted in numbers, for instance 0 equal NO and 1 equal YES.

	Age	Smoker	Familiarity	Daily Sport Hrs	Tumour
<i>Mario Rossi</i>	58	Yes	Yes	1	Yes
<i>Anna Bianchi</i>	49	No	No	3	Yes
<i>Piero Verdi</i>	73	No	Yes	1.5	No

Table 2: example of a table for Supervised Learning

The databases usually contains a huge amount of records, and it is split into two pieces, called **training set** and **test set**, usually with the first larger than the second: the first one will be the very “literature” used to train the algorithm in the **training phase**, while the second will be the dataset used to understand if the learning process is gone well or not in the **test phase**.

Then the last thing to decide before really starting learning, is choosing the mathematical framework that must be followed to generate the model. Usually, these approaches work through **iterative optimization** of an **objective function**, so they need several repetitions before reaching an acceptable result.

The learning process aim is to create a model that maps the different records of the training set in input into an output (the target) as a function of the values of attributes.

In other words, the training phase generate a function that assigns to a record in a dataset one of the possible target values, as a function of its attribute's values: this behaviour is called **prediction**, so the ability of an algorithm to assign to an unlabelled record a target value.

After the training phase, the next one is the test phase: now the model is fed with the records of the test set, so a table with the same structure of the training one (same attributes) but robbed of the target values. Now the model must predict the target of the new records by itself: if the accuracy of the process satisfies the expectations, the model can be then implemented in the future; conversely it must be done again the training phase. For *accuracy*, is intended the percentage of records from the test set that are being correctly predicted.

A huge number of approaches are available in the SL problems: *linear regression*, *decision trees*, *support vector machines* are just examples of algorithms. There are not *best approaches*, because there are not fixed parameters that work always well. For these reasons, usually more than one method is used, so it is possible to choose the approach that gave the best result on the dataset used.

A question that could be ask at this point would be: what is an acceptable accuracy? The answer is not straightforward: indeed, there is not a magic value that decrees if an algorithm is working well or not but depend on trade-off between **discrimination** and **generalization**: the first refers to the ability of the system to recognise which are the most significant attributes that generate the target, while the latter is the prediction accuracy using the same attributes.

Since in the learning process a specific database is used, if the accuracy is too high, we are not teaching “how to predict the target of a record in the situation we are studying”, but the system is learning perfectly the patterns and the correlations that are hidden inside “the specific database we are using in training”. In other words, the system did not understand which attributes are significant for the phenomenon under analysis, but those to perfectly replicate the training set used. So, while the accuracy of the learning process becomes higher, in the test phase it will decrease a lot. The trade-off is to accept a reasonable number of errors, that let the machine being less accurate in the training phase (so loosing discrimination ability), but more precise in

the prediction of new records, because it is more freely to exploit all the attributes that it considers important.

### Example: Support Vector Machine

As example of Supervised Learning the Support Vector Machine (SVM) model will be briefly introduced. This model can be used for both regression and classification with the relative adjustment, and it is one of the most robust method [12] available.

Two sets of points are called *linearly separable* if there exists a hyperplane that separate them in the space of dimension  $\mathbb{R}^n$  [10]; in a bidimensional case (target has 2 states) the hyperplanes reduce to lines as shown in Figure 3: an infinite number of line exists, but only the one *equally distant from the nearest points of the two dataset*, so the one with the most generalization ability, will be chosen as the *separating line*.

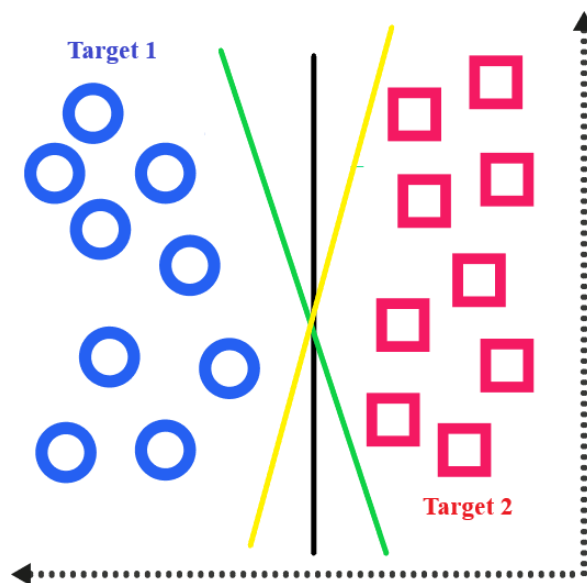


Figure 3: example of the starting point of a SVM learning methods

The nearest points of the two datasets are considered and an infinite number of lines passes through each one of these points: the system will choose the two parallel ones that creates the maximum *margin of separation* between the two points, and they are

called *canonical supporting hyperplanes*. Then from all the possible hyperplane, the algorithm chooses the one parallel to the canonical ones, that minimize the distance between itself and the two canonical margins: the only possible solution is the line that separate the margin exactly at the middle, so the distances between the *separating hyperplane* and the two nearest points (called *support vectors*) are the same.

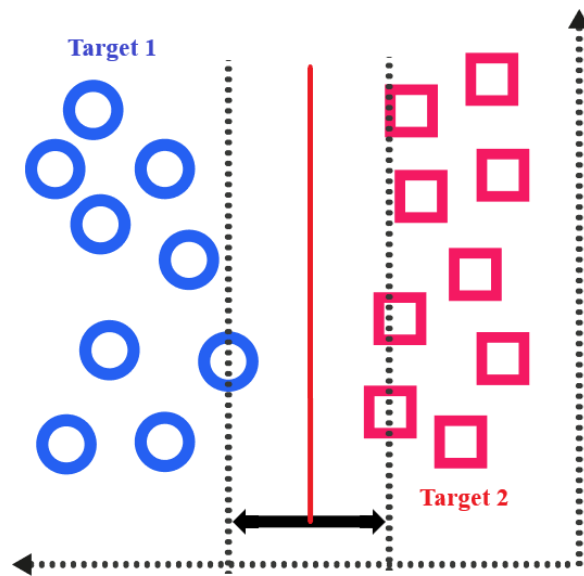


Figure 4: example of separating hyperplane in the SVM process

The hyperplane of separation, like the one in Figure 4 will be used, after the training phase, as a *decision function* for the prediction process. The example presented is of course an ideal one: the points are *linearly separable*, so a straight line is sufficient to separate the two sets, but the SVM also includes the possibility to apply nonlinear functions, as shown in Figure 5.

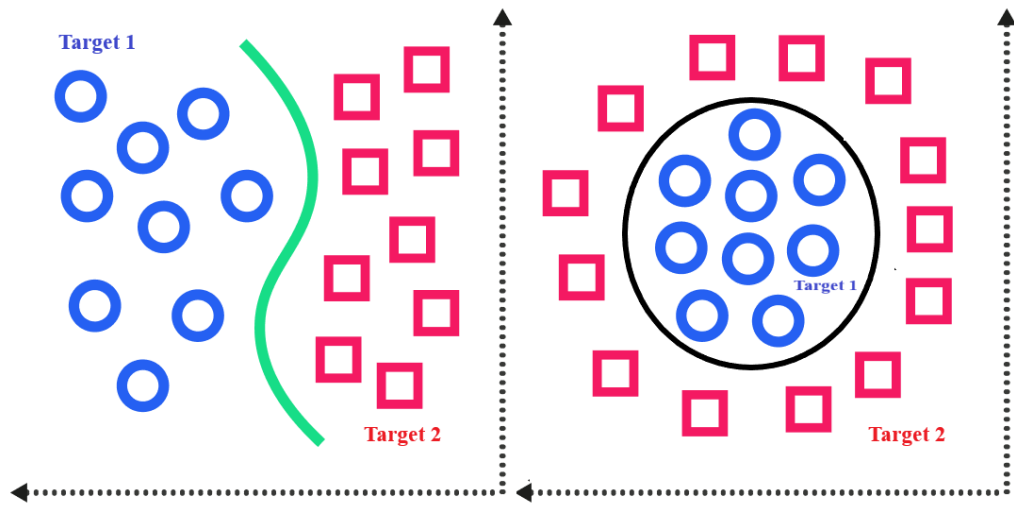


Figure 5: examples of nonlinear separation in SVM

### 2.1.2 Unsupervised Learning

Under the UL methods, are grouped all those algorithms that does not require labelled data in the learning process, but they analyse the information given to them in order to capture patterns or probability density hidden in the datasets by themselves [13]. The main applications are *clustering* and *anomaly detection*: the first has the aim to create different clusters (groups) in which records are in some sense similar to each other, respect to the ones in the other groups; in the latter the system try to identify the rare event or record in the dataset, so that has totally different characteristics than the other records.

The UL is used for different purpose and finds application in the most disparate field like psychology, physiology, image recognition and so on. Clustering is able to identify some homogeneous groups inside a population, and for instance is highly used in marketing applications to understand the different behaviours of people, depending on their nationality, age, salary and so on. But also, into clinical trials, to create population of patients, for instance, with the same response to a medical situation, such as the administration of a new drug. Anomaly detection is applied particularly in bank frauds or algorithms that ensure security of websites and e-commerce sites, due to his ability to discover strange behaviours of sellers or buyers.

Clustering and Anomaly Detection exploits similar techniques to reach their goal, as both try to separate groups of records from each other or one record from all others: for this reason, in the next pages, only clustering will be further discussed.

As indicated in [10], clustering methods must fulfil some general requirements like:

- Flexibility: some clustering methods can be applied only on numerical records to exploit Euclidean metrics, but a flexible approach to use also categorical values are required.
- Robustness: the algorithm must be stable, so the cluster assigned to a record must not change, with respect to small changes in its attribute's values. This ensure that the method is not affected by noise in the data. Another sign of robustness is the stability with respect to the variation in the presentation order of the same records to the system.
- Efficiency: usually datasets have huge dimensions, so the system must be able to achieve the goal in a reasonable computing time. This, of course, decrease the robustness and the accuracy of the method, so it is necessary to find a trade-off between all these characteristics.

Different approaches are available and in the next pages an example of one of these algorithms will be presented: the *K-means algorithm*.

### Example: K-means algorithm

The algorithm works trying to group the records as function of their distance from some points chosen as *centroids* of each cluster. So, the system receives in input a dataset  $D$ , the number of clusters  $K$  to be generated and the definition of a distance function  $dist(x_i, x_k)$ , for example the Euclidean distance:

$$\begin{aligned} & \text{for two point } P = (p_1, p_2) \text{ and } Q = (q_1, q_2) \text{ the distance } dist(p, q) \\ & = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2} \end{aligned}$$

Before starting, during the *initialization phase*, a  $K$  number of records are chosen randomly from the dataset as the centroids of clusters. In the first iteration, the algorithm calculates the distance between all the centroids and each record, and each one is assigned to the group that has *the most similar centroid*, which means the *nearest one*.

After all the records are assigned, there are two possible scenarios: the last iteration did not change the groups, so all the records are assigned to the nearest centroid, and the algorithm stops. Conversely, the new centroid is computed as the mean value of all the records that belong to the cluster, and then it is recalculates the new assignments in a further iteration.

An example is given graphically in Figure 6, with the clusters identified by blue and red colours and the centroids as starburst.

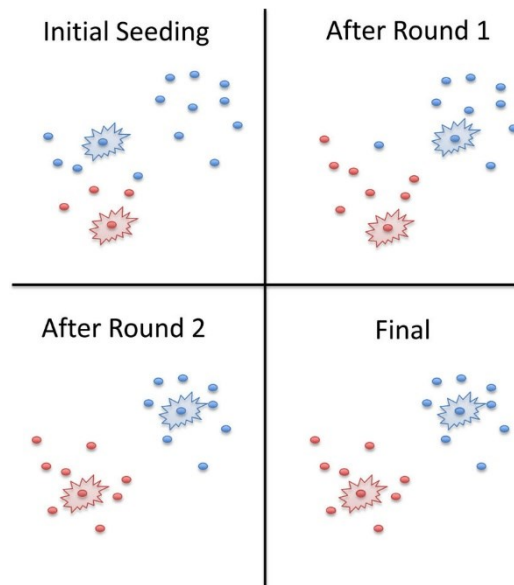


Figure 6: example of application of the K-means algorithm

Starting from top left, the centroids does not separate properly the records, as some points are closer to the red than the blue one: the algorithm calculates the new centroids as the mean of the values and computes the new clusters. Then a second iteration is necessary to reach the perfect separation and the final result is given in the bottom right.

In this paragraph the concept of ML was introduced and analysed to understand which are the main approaches, and which are the different possible applications of this kind of AI.

In conclusion, the ML has great potential, and is already now spread in a lot of automatic systems and products. But in the recent decade, the rapid technological development seems to have left this approach behind. Indeed, a couple of problems slowly started to arise.

The ML algorithms work well when data is structured into tables, with well-defined attributes and values, while fail when data is in form of images, videos, or sounds. Depending on the phenomenon under analysis, it is not straightforward to build a table with the parameters that could be significant: for instance, in the situation of face recognition, it is really hard to build a structure in which the attributes are “nose”, “eyes” or “mouth”, and then assigning values that have meanings. So, the first problem is that if I want to exploit different kind of data, I should not use ML.

Another difficulty when using this method rises from the necessity of a constant intervention of humans. When introducing ML, we said that “ML refers to the capability of computers to learn without being explicitly programmed”: as we said in the previous pages, the algorithms do not need to be “programmed” in the traditional way, but it still requires that the programmer decide the model to use in the learning process. Sometimes different frameworks require a specific structure of the dataset so there must be someone who manipulates the information, so it is like he is writing the code to rule the system behaviour. This is not a real problem, but it is possible to discuss the fact that the ML analysis is highly affected from how humans approach the world. So, the second problem is: is there a system that can independently learn from the data, with less human intervention as possible?

The solution to overcome these problems is offered by the Deep Learning approach. In the next paragraph it will be presented and compared to the ML method, trying to understand which are the improvements that this technology brought and why you should prefer the DL approach.



## 2.2 Deep Learning

The Deep Learning approach can be considered as a subset of the ML, as it also learns from data and extrapolate some characteristics hidden in the dataset. However, the term *deep* refers to the ability of these systems, to analyse and understand the data in a way similar to how the human brain works: deeper means “a multi-level learning” [14], so the ability to see something not only as a stand-alone thing, but understanding the different facets that composed what we are focused on.

For instance, if someone tells us to think of an apple, in our mind does not exist a unique image for that fruit. We can imagine an infinite number of apples, because we learnt that an apple has different shapes, different colours, different dimensions. Our brain, if asked to generate the image of an apple, activate a process of going from a bottom layer to a top one, with an increase of generality going from the first to the last step: the bottom layer may contains the basic information like contours and dimensions in 2D, then in the upper layer it may generate the tridimensionality, and then another level could be the colour, and finally the last one in which it adds the last details.

So, the main difference between ML and DL is that the latter has the powerful ability to treat the information of the learning process to generate as output an abstract concept, not a specific one. In the ML approach, the final output was a sort of unique information, that can assign a defined label (classification) or creates specific populations (clustering) based on data it processed. In DL, the output can be an abstract concept of something, so it learns from the data which are the characteristics of *an apple*, and not the ones of *that apples described in the dataset*.

The other side of the same coin is that to reach this level of knowledge, the DL approach needs more information than the ML, and of course the computational time is longer. The biggest problem, that initially undermined the development of DL systems, is the great power required for hardware resources: indeed, the first DL projects date back to the 40s and 50s, but they found a complete development and spread of applications just in the last decade. In fact, only the advent of GPUs in the 2010s, that significantly improve their computational power, make it possible to train the DL systems efficiently [15].

In the previous paragraph, the main limitations of ML were listed. One of these was that these kinds of algorithms can treat mainly information structured in tables: DL methods can overcome this, as they are able to exploit not only structured data like tables, but also *raw data* without manipulation like images, videos, and audios. This ability has made the DL approach one of the most powerful tools developed in the last decades. For instance, if we want to train a DL method to recognise an apple, we can use a database filled of thousands of images of apples: how to build such a system will be later explained, but for now is sufficient to say that the learning process in DL is influenced only from the structure of the system, and no human intervention is needed in the learning process.

Indeed, the possibility to feed a DL algorithm with information coming directly from the world, overcome the necessity of human intervention in manipulating the datasets. So, there is no need for a programmer to extrapolate features from dataset in order to be usable by the machine, but just to collect as much information as possible.

So, the main differences between the ML and DL approaches are the structure of the systems, the lower need for human intervention in the DL and the larger data that the latter requires [16].

The advent of DL brought with it an era of innovations, in which AI has settled in the most varied fields of application, leading to the development of technologies that had remained only theoretical for several years, like self-driving cars, face recognition and elaboration of human language.

The main DL structure is the Artificial Neural Network (ANN), which are computing systems inspired by the biological neural networks that constitute animal brains [17]. In the next pages this structure will be described, followed by the different topologies required depending on the application, and a couple of examples of their utilization.

### 2.2.1 From Artificial Neuron to Artificial Neural Network

The ANN seems to be an invention of the 2000s, but the reality is that the development of this kind of systems begins in the '40s. McColluch and Pitts in 1943, laid the theoretical foundation for what Frank Rosenblatt would create in 1950s: the ancestor of the actual DL systems, the Perceptron. This machine was built for image recognition, and contains photocells as receivers, connected to multiple neurons which classify the inputs created by the photocells [18]. From that starting point began the development of ANNs that, after 70 years, will evolve into deep learning.

An ANN tries to mimic the structure of the brain, whose cells are called *neurons*, the smallest working units of the brain. These units are grouped into networks called *neural circuits*, so a population of neurons interconnected through synapses that carry out a specific function when activated [19]. The structure of the neuron is shown in Figure 7: the input signals coming from other neurons are caught by the *dendrites*, that route these signals to the cell body called *soma*. Here the information is elaborated and can activate or not the neuron response; if the elaboration process activates the cell, an output signal called *action potential* is generated and sent through the *axon* until it arrives to the terminal regions of the latter. Now the signal can reach another neuron through a chemical or electrical phenomenon, depending on whether the distance between the two adjacent cells is large (chemical synapse) or small (electrical synapse).

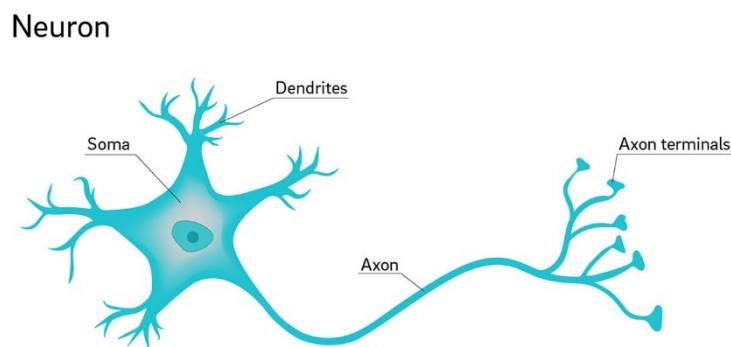


Figure 7: example of a human neuron

The artificial neuron is a mathematical model that mimics the biological neuron: it receives **one or more inputs**, processes these inputs as an **integration** of them through some **weights** and finally **activates** itself if this integration **overcome a threshold**. The output of each neuron is called **state of the neuron** and depends on an **activation function** that determines the value of the output.

The artificial neuron replicates this behaviour as illustrated in Figure 8, where a neuron (depicted as the circle in the centre) receives the weighted inputs like in the dendrites, integrates these signals like what happens in the soma and, if the sum overcomes the threshold (called bias in the figure), it becomes active, and the output is governed by the activation function, namely the elaboration process that happens in the biological cell.

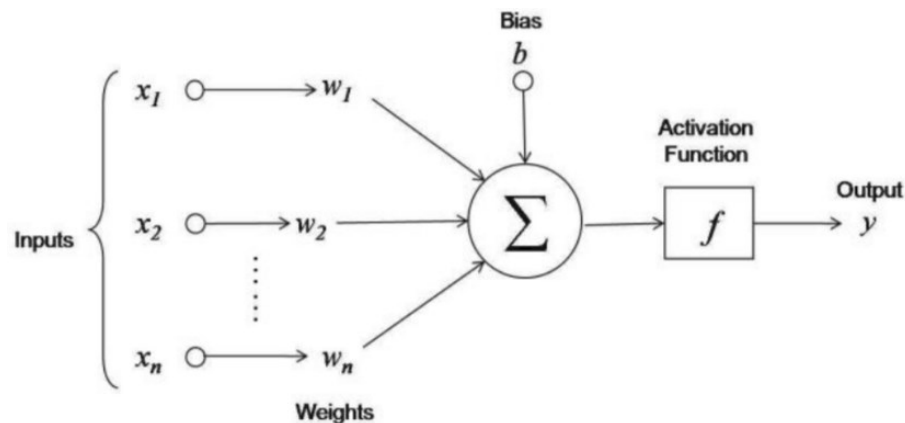


Figure 8: example of neuron principle

Both in the biological and artificial case, the inputs of a neuron are the output signals coming from other cells: in nature the different neurons are connected through synapses, and each one of these is characterized by a *strength*, therefore a minor or greater influence on the cell to which it is connected. In the artificial representation, each input is characterized by its strength, which is depicted as a *weight* that can excite (positive value), inhibit (negative one) or even be ineffective (weight equals zero) to the artificial neuron. As the neuron sums up the inputs (just as example, we now consider only positive inputs), a positive weight increases the sum, while a negative one decreases the integration. Sometimes neurons can also have *self-*

connections, so the output of a cell becomes an input for itself: in this case, when the neuron is activated, it is able to excites or inhibits itself depending on the positive or negative value of the weight of the self-connection.

The process that generates the output from the inputs starts with the weighted sum. Exploiting the structure of Figure 8, let us call  $P$  the sum of the inputs  $x_j$  weighted by  $w_j$ . So, the inputs are summed up into the value  $P$  in the soma (the circle in the middle of Figure 8), then this value becomes the input of the activation function  $f$ , after the subtraction of the threshold  $T$  (bias in Figure 8), and generates the output  $y$ :

$$P = \sum_1^n x_j w_j \quad \text{then} \quad y = f(P - T)$$

The activation function characterizes the behaviour of the neuron, that can completely changes depending on the function chosen. There are multiple activation functions available: in case of binary output (0/1 or -1/1), the **Heaviside** and **signum functions** are usually used; while extending the output to any real value, linear and non-linear functions can be chosen, like the **logsig** or the **tanh** function.

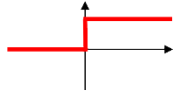
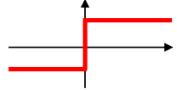
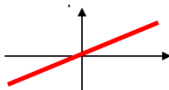

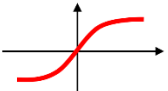
Activation function	Equation	1D Graph
Unit step (Heaviside)	$\phi(z) = \begin{cases} 0, & z < 0, \\ 1, & z > 0, \end{cases}$	
Sign (Signum)	$\phi(z) = \begin{cases} -1, & z < 0, \\ 0, & z = 0, \\ 1, & z > 0, \end{cases}$	
Linear	$\phi(z) = z$	
Logistic (sigmoid)	$\phi(z) = \frac{1}{1 + e^{-z}}$	
Hyperbolic tangent	$\phi(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$	

Figure 9: examples of activation functions

In Figure 9 it can be noticed that the Heaviside and Sign functions can assume only two values, so they can be used only in binary problems (actually  $\text{sign}(z) = -1$  if  $z = 0$ ). The others can be applied if the output can assume real values: the main difference is that in the Linear function the output can assume **all real values**, while in the Logistic sigmoid (logsig) and Hyperbolic tangent (tanh) the output is bounded between 0 and +1, and -1 and +1 respectively.

Now is it possible to understand the meaning of the threshold: in our neuron, in order to generate an action potential (so the generation of an output in response to a stimulus), the input stimulus must exceed a certain voltage potential, under which the cell does not provide any response. The same happens in the artificial neuron: in the situation of a binary output, for instance 0 and +1, it is possible to associate the 0 value to an “idle state” and the 1 to an “active state”; if the artificial neuron’s threshold is, for instance, 1, the inputs of the artificial neuron must overcome this value to generate a response. Indeed, if the inputs sum is less or equal to the threshold, the neuron does not provide any response, as in the biological one if the electrical stimulus does not overcome the bias value; when the sum exceeds the threshold, the neuron generates a positive output (so an excitatory one) similar to the action potential.

The artificial neuron model can be used, for instance, in taking a decision as output, getting as inputs different circumstances that can affect the final decision. For instance, I cannot decide “if I want to go for a walk to the beach or not”: we can assign the -1 value to the decision “I do not want to go” and the value +1 to “I want to go”. The inputs are the situation that influence the decision, and to each one is possible to assign a positive value if it supports the “to go” choice or a negative value at contrary: an input can be for example “if the weather is good or not” with a positive value (+1) for good meteorological conditions and a negative (-1) for the contrary. Then we must assign a weight to all the inputs based on how decisive that aspect is on the final decision, for instance in a range from 0 to 1.

Aspect	Weight	Answer
Is the weather good?	1	YES = +1

		NO = -1
Am I tired?	0.5	NO = +1
		YES = -1
Do I prefer to play football?	0.2	NO = +1
		YES = -1

Table 3: example of which could be the determining aspects that affect the final decision

In Table 3 are summarized some possible inputs that can determine the final decision, while in Figure 10 it is depicted the structure of the neuron that can decide to go for a walk or not. The three inputs can assume the values -1 or 1, while in circles are shown the weight values and in the top the threshold value (in this case it is 0.5). For instance, let us assume that today the weather is good, and I do not prefer to play football, but I am a little bit tired: so, following the values in Table 3, the first and third input is +1, while the second is -1. The activation function is the Signum one, so the output can be -1 if the input is less or equal 0, +1 if the input is higher.

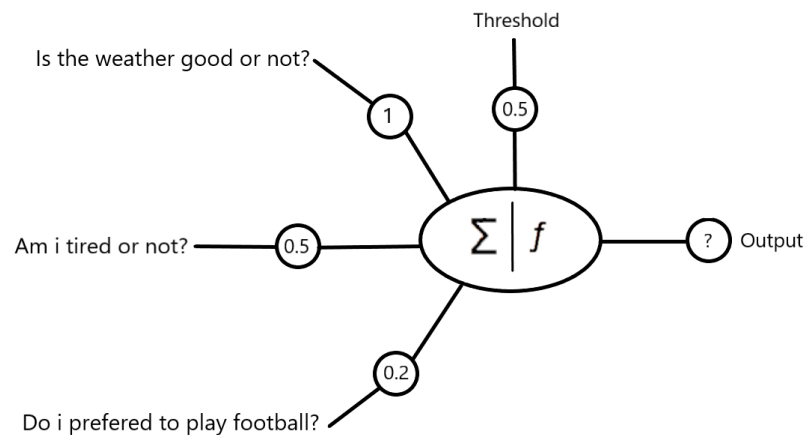


Figure 10: structure of an artificial neuron that can decide to go for a walk or not

The process starts with the weighted sum, so:

$$P = \sum_1^3 x_j w_j = (+1 \cdot 1) + (-1 \cdot 0.5) + (+1 \cdot 0.2) = 0.7$$

Then the activation function is applied, taking as input the sum  $P$  to which the threshold is subtracted:

$$y = \text{sign}(P - T) = \text{sign}(0.7 - 0.5) = \text{sign}(0.2) = +1$$

So, the neuron decides that I have to go for a walk, but for example if the threshold was higher, for instance 1, the model would give a different result: the activation function would take as input a value of  $-0.3$ , giving a negative result.

Now that the description of the single neuron is clear, it is possible to shift the attention to the structure that underlies the deep learning approach, the Artificial Neural Network.

Imagine you want to apply the previous model to something closer to a technical application, for instance a neuron able to replicate the behaviour of a Boolean port, like the AND port. This port takes two inputs, whose values can be 1 or 0, and gives an output equal to 1 only if both the inputs are 1. In Figure 11 is depicted the input/output relationship of an AND port, and the output position inside a space built on the two inputs values.

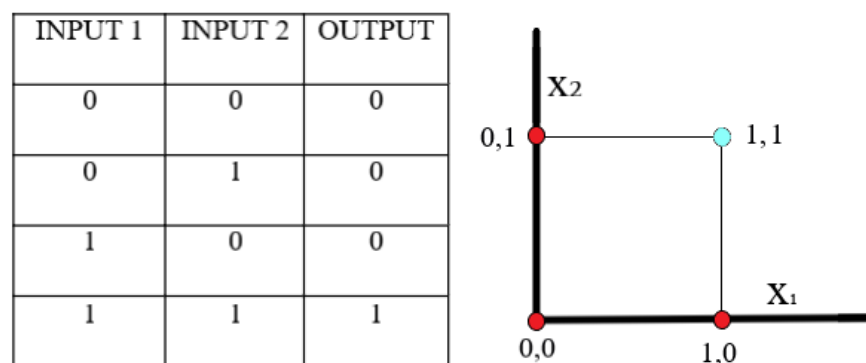


Figure 11: I/O relationship of an AND port and visualization of the output in the inputs space.



So, our neuron must be able to separate the (1,1) point from the other three points: this can be done assigning the same weight equal to 1 to the inputs, a threshold value equal to 1.5, and exploiting the signum activation function, as shown in Figure 12.

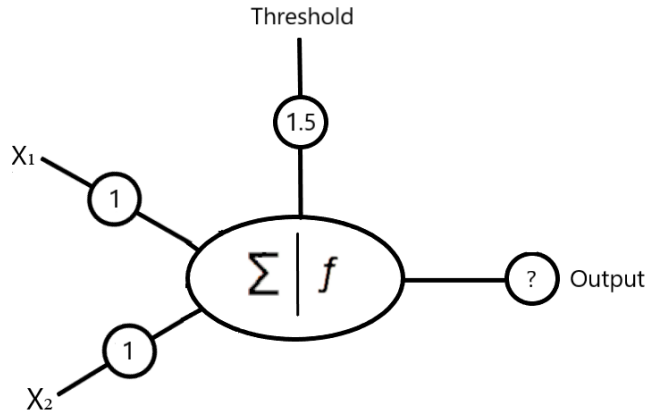


Figure 12: structure of an artificial neuron as an AND port

With this configuration the neuron can solve the problem, separating the points in the correct way:

$$\text{for } x_1 = 0 \text{ and } x_2 = 0 \quad ; \quad y = \text{sign}(0 \cdot 1 + 0 \cdot 1 - 1.5) = \text{sign}(-1.5) = -1$$

$$\text{for } x_1 = 1 \text{ and } x_2 = 0 \quad ; \quad y = \text{sign}(1 \cdot 1 + 0 \cdot 1 - 1.5) = \text{sign}(-0.5) = -1$$

$$\text{for } x_1 = 0 \text{ and } x_2 = 1 \quad ; \quad y = \text{sign}(0 \cdot 1 + 1 \cdot 1 - 1.5) = \text{sign}(-0.5) = -1$$

$$\text{for } x_1 = 1 \text{ and } x_2 = 1 \quad ; \quad y = \text{sign}(1 \cdot 1 + 1 \cdot 1 - 1.5) = \text{sign}(+0.5) = +1$$

What the neuron is doing, is creating a separation line between the points, generating two hyperplanes that contain points of only one value. This result is depicted in Figure 13.

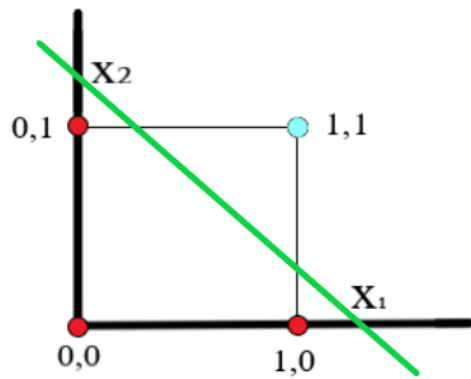


Figure 13: separation line between points

The same process can be done also to mimic the behaviour of other Boolean ports, but not all of them: for instance, it is not possible to replicate the XOR (Exclusive OR) port: this Boolean operator, takes  $N$  inputs and gives a TRUE (1) output only if the number of TRUE inputs is odd. For example, with two inputs the XOR port gives an output equal to 1 when the two inputs are different from each other.

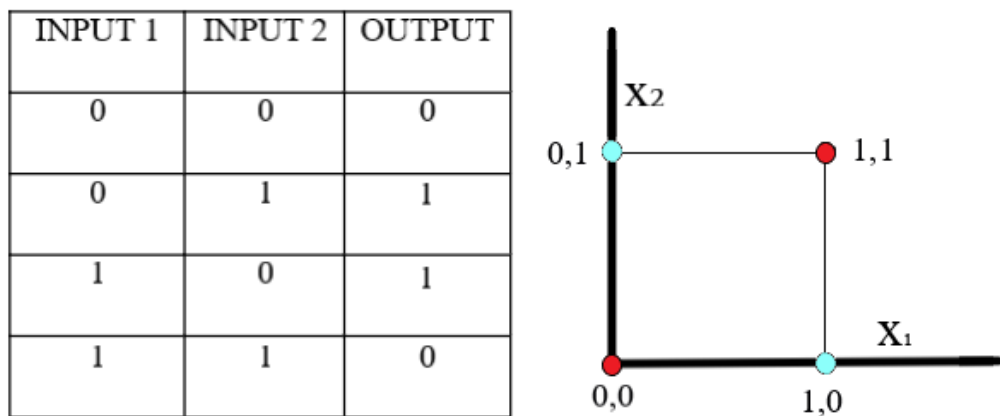


Figure 14: I/O relationship of an AND port and visualization of the output in the inputs space

From Figure 14, it is possible to understand that a single separation line is not sufficient to separate the points. This also means that a single neuron is not enough, and we need to add other ones in order to complete the task.

Indeed, as it can be seen from Figure 15, the XOR port can be seen as the joint between the OR port and the negation of the AND of the two inputs [20], in a process that requires three steps: firstly the two inputs enters in two neuron that perform the AND and OR task; then the outputs of these neurons become the inputs of two new neurons, one perform the negation of the AND, while the other is just an identity neuron, that let the input signal pass unaffected. Finally, the last neuron performs an AND operation, taking as inputs the outputs of the previous neurons.

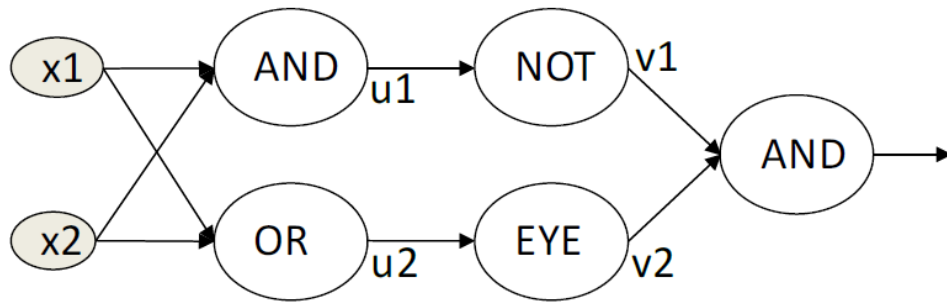


Figure 15: topology of the network that perform the XOR operation

An artificial neural network has just been created, and the process that produce the output  $y$  from the inputs is presenting exploiting two examples, one that takes as input the couple  $(1,0)$ , and the other the couple  $(1,1)$ . In the example, the weights will be considered all equal to 1, this is the reason of their absence in the picture. The AND operator is represented by the product symbol, the OR operator is depicted as the sum and the NOT with the bar above the number ( $\bar{\phantom{x}}$ ):

$$\begin{array}{l} \text{for } x_1, x_2 = (1,0) \quad u_1 = 1 * 0 = 0; \quad v_1 = \overline{u_1} = 1; \\ \quad \quad \quad \quad \quad \quad u_2 = 1 + 0 = 1; \quad v_2 = u_2 = 1; \quad y = v_1 * v_2 = 1 * 1 = 1 \end{array}$$

$$\begin{array}{l} \text{for } x_1, x_2 = (1,1) \quad u_1 = 1 * 1 = 1; \quad v_1 = \overline{u_1} = 0; \\ \quad \quad \quad \quad \quad \quad u_2 = 1 + 1 = 1; \quad v_2 = u_2 = 1; \quad y = v_1 * v_2 = 1 * 0 = 0 \end{array}$$

The structure in Figure 15 is just an example of how a network can be built, but it shows one of the main characteristics of the artificial neural networks: the structure is composed by different **layers**. The one that receives the inputs is called **input layer**, while the one that produce the output is the **output layer**; between these two, there can be zero or multiple layers called **hidden layers** (in the example above, there is one hidden layer).

Between two layers, different patterns of connection can be present [17]: the neurons can be **fully connected**, if in one layer all the neurons are connected to the ones in the next layer; or they can be **pooling**, so a group of neurons in a layer are connected to only one in the next layer.

Finally, another characterization of an ANN is given by the *direction of propagation* of the signals: if they proceed only from the input to the output through the different layer, it is a **feedforward neural network** (FFNN); while if there are some connections between a layer and the same or the previous one, the network becomes a **recurrent** one.

ANN can have several layers of different numbers of neurons, and they are called *deep* neural network, like the one shown in Figure 16: it is a feedforward network, as the signals can only proceed in one direction, and fully connected, as all the neurons of a layer are connected to the ones in the next layer. At the beginning of this chapter it was said that the term *deep* refers to a “multi-level learning”, and this is represented by the different layers in the network depicted: still retaining a general view of the problem which will be addressed in more detail later, each layer of the network is able to extract parameters of the data in a more abstract way as the signals proceed from one layer to the next one.

Using the example given before of “thinking of an apple”, the network works understanding which are the features of an apple from the simplest ones in the first layer, like the contours, and then proceeding in the next layer where the dimensions are learnt, and then moving to colours and so on.

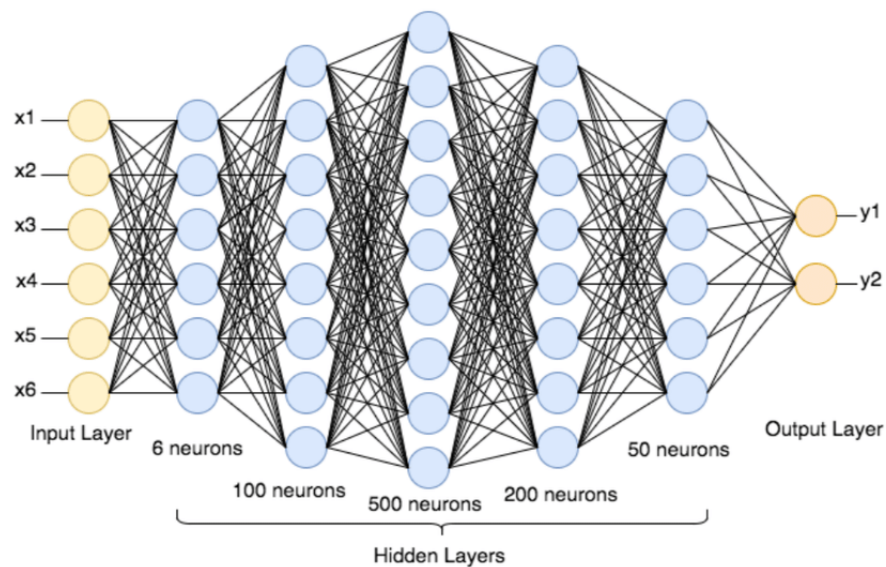


Figure 16: example of a deep neural network

Due to their huge learning potential, ANN find a lot of different application: from simple tasks like the Boolean reproduction seen before or the approximation of functions, to really difficult problems, like the image recognition, the language interpretation and other complicated task. The strength of the ANN approach is given by the learning procedure, that indeed is called *Deep*.

In the DL approach, the mathematical model is the artificial neuron and the ANN. In this paradigm, learning means to specialize the network to a specific task [20]: given the topology of the network (so its structure), the ability to realize a task is given by a specific combination of the weights and threshold, and to define them it is necessary to **train** the network.

In a way similar to the ML approach, there are different way to train an ANN, for instance the Supervised Learning is available also for these algorithms: as shown in the 2.1 Machine **Learning** chapter, this approach consists in feeding the algorithm (the network in DL) with some example inputs of which we already know the output, through a mathematical model the system learns which are the correlations between the output and the inputs, and finally the algorithm is tested with other known records.

But the DL and the ML approaches are different due to the mathematical frameworks on which they are based: in ML the output was given by an optimization process of an objective function that ruled the behaviour of the system, and the analysis of the records allowed the regulation of the function parameters in order to obtain a good accuracy in the prediction. These parameters can be directly corrected through the optimization process because the function maps directly the output with respect of the input, so the entire training set used to feed the learning process.

In ANNs, the structure of the networks makes the definition of the parameters (weights and threshold) not easy to implement; in particular, as the number of layer increases, it becomes difficult to correlate the output error with the different neurons that belong to deep hidden layer. In the following pages, the training technique for ANNs will be described, as a fundamental knowledge to understand the deep learning approach.

## 2.2.2 Training an ANN

To comprehend the training process of a network, it is essential to first understand what the learning process of a single neuron is: the *perceptron learning rule* is used to train a single neuron. It starts from a vector of weights  $w_{start}$  arbitrary defined and, through an iterative procedure which exploits the difference between the target output and network one, it arrives at a solution  $w^*$ .

Let us consider a neuron that take  $M$  input and  $N$  output, with a signum activation function, so the output can be only  $\{-1; +1\}$ . In order to define and correct also the threshold as one of the parameters of the system, it will be considered from now on as an added weight  $S$  of a virtual input always equal to  $-1$ . The training set is composed of  $R$  examples, of which are known the desired target. The perceptron rule is the following:

$$w^{(k+1)} = w^{(k)} + \Delta w^{(k)} \quad \text{with} \quad \Delta w^{(k)} = \eta(t^{(k)} - u^{(k)})x^{(k)}$$

Where  $k$  is the actual record passing through the neuron;  $t^{(k)}$  is the target of the actual record,  $u^{(k)}$  is the actual output of the neuron and they can be  $\{-1, +1\}$ .

$x^{(k)} = [x_1, x_2, \dots, x_n]^{(k)}$  is the input vector and  $w^{(k)} = [w_1, w_2, \dots, w_n, S]^{(k)}$  the weight vector.  $\eta$  is a parameter called *learning rate*, positive and less than 1, that define the influence of each correction on the parameters of the network.

The meaning of the rule is easier to understand, if you consider what the neuron is actually doing: as shown in the example of the Boolean operator at page 24, the neuron tries to implement a separation line that generate two hyperplanes, each one containing only records belonging to one target class. Considering that this separation line is generated through the weights vector, the perceptron rule, at each iteration, reorients this vector toward itself if the error signal is +1 or opposite to itself if the error signal is -1, while it does not provide a correction if the error is equal to 0 [20].

$$\begin{aligned} \text{If } e = +1 \quad \text{then} \quad w^{new} &= w^{old} + x^{(k)} \\ \text{If } e = -1 \quad \text{then} \quad w^{new} &= w^{old} - x^{(k)} \\ \text{If } e = 0 \quad \text{then} \quad w^{new} &= w^{old} \end{aligned}$$

So, it can be said that each pattern contributes to a direct correction of the weights, but only if there is a misclassification (so an error), otherwise nothing will be done.

This learning approach suffers from the typical problems of these kind of algorithms: one is the existence of the solution and the other one is the convergence of the algorithm to the desired solution, if this exists. In this case, with a single neuron, the solution exists only if the training set is *linearly separable*. This is a very strict constraint, as in practice, it is hard to find such a dataset, so it is necessary to extend the learning process to neuron that exploit continuous activation functions, and then increase the number of layers in the ANN.

So, let now consider the same neuron as before, with  $M$  input and  $N$  output, a training set of  $R$  records, but with a sigmoid activation function; the overall error signal can be computed as the sum of the error signals for each pattern  $k$  of the dataset, with the latter calculated as the squared value of the difference between the desired target and the actual output of the network:

$$E^{(k)} = \sum_{k=1}^R E^{(k)} = \sum_{k=1}^R \sum_{i=1}^N \frac{1}{2} (t_i^{(k)} - u_i^{(k)})^2$$

It can be expected that the correction of the weights vector  $\Delta w_{ij}$  is a function of this error, so that  $\Delta w_{ij} = f(E)$ . In particular, the delta rule assumes the correction  $\Delta w_{ij}$  must be negative if  $E$  increases with the increase of  $w_{ij}$ , while it must be positive if  $E$  increases with the decrease of  $w_{ij}$ .

To reach this behaviour, the gradient descending method is used: it is an iterative optimization that tries to reach a local minimum of a function, taking steps in one direction or the other one, depending on the negative of the gradient calculated on the actual point. So, it computes the gradient of the cost function  $E$  with respect to the weights value, and change each weight in the negative (opposite) direction to the gradient:

$$w_{new} = w_{old} + \Delta w \quad \text{and} \quad \Delta w = \eta \left( -\frac{\partial E}{\partial w} \right)$$

So, the updating rule of the weight for a neuron  $i$  with an input  $j$ , given the  $k$  record of the dataset, becomes:

$$(\Delta w_{ij})^{(k)} = \eta \left( t_i^{(k)} - u_i^{(k)} \right) f' \left( P_i^{(k)} \right) x_j^{(k)}$$

Where  $f' \left( P_i^{(k)} \right)$  is the derivative of the activation function of neuron  $i$ , in correspondence of the summation  $P$  at record  $K$ .

So, the algorithm work to reach the convergence to a solution, until a *stopping rule* terminates the procedure: this rule can be the maximum number of iterations, an error threshold below which the result is considered acceptable, a gradient too small or combination of them.

But even the Delta rule has some limitations: the computation of the derivative implies constraints in the choice of the activation function, that must be differentiable. Another problem is the choice of the initial points of the gradient descending, which suffers from local minima issues. The most important limitation is that the delta rule cannot be applied to networks that present hidden layer: this because there are not target for the output of the neuron belonging to these layers, so there is no correction available.



The solution of this problem is offered by the **back propagation algorithm**, which is a generalization of the delta rule: this approach is again an iterative process, that tries to back propagate the error signal from the output back to the hidden layer in a sort of “reverse direction of the signals”. The process is divided into two steps, one that follows the classic direction and permits to calculate the output from the actual input, and the second that back propagates the computed error through the hidden layers of the network.

To describe the algorithm, the example in [20] is follow: it considers a multi-layer Perceptron with 4 inputs, one hidden layer of three neurons and an output layer composed by two neurons, as depicted in Figure 17.

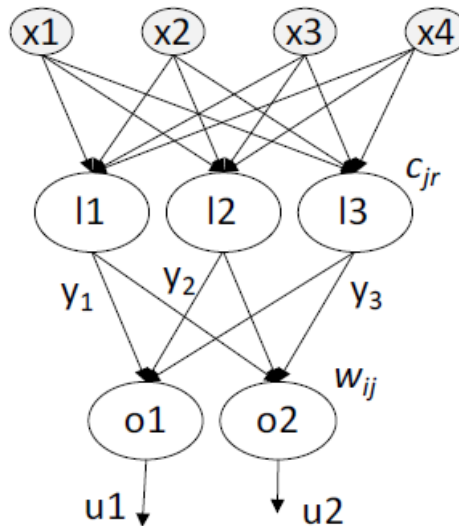


Figure 17: example of perceptron with a hidden layer

Being in a supervised learning approach, it is possible to compute the correction  $\Delta w_{ij}$  using the delta rule, exploiting the target value and the output of the system; the same could be done for the output of the hidden layer, but no targets are available for these neurons. The backpropagation algorithm overcome this limitation:

$$\Delta w_{ij} = \eta \sum_{k=1}^R \left( t_i^{(k)} - u_i^{(k)} \right) f' \left( P_i^{(k)} \right) y_j^{(k)}$$

$$\Delta c_{jr} = \eta \sum_{k=1}^R (\bar{y}_j^{(k)} - y_j^{(k)}) f'(P_j^{(k)}) x_r^{(k)}$$

Where the first equation is the classic delta rule applied on the output, as if the hidden layer were the input layer, while the second one is the delta rule applied to the output of the hidden layer in respect to the input from the input layer.

Considering a single pattern in the training set, and according to the gradient descent principle, the gradient of the error  $E$  with respect to the weight of the hidden layer  $c_{jr}$  is computed as:

$$\Delta c_{jr} = f \left( \frac{\partial E}{\partial c_{jr}} \right)$$

By chaining the differential with the factor  $\partial P_j^H$  namely the differential of the action potential of the hidden neuron  $j$  (H superscript stands for “hidden”), and then chain to the factor  $\partial y_j$  (output of the hidden neuron  $j$ ) it can be expressed as:

$$\frac{\partial E}{\partial c_{jr}} = \frac{\partial E}{\partial P_j^H} \frac{\partial P_j^H}{\partial c_{jr}} = \frac{\partial E}{\partial y_j} \frac{\partial y_j}{\partial P_j^H} \frac{\partial P_j^H}{\partial c_{jr}}$$

The term  $\frac{\partial E}{\partial y_j}$  can be chained to  $\partial P_i^O$  which represent the differential of the action potential of the neuron  $i$  in the output layer. However, chaining  $\partial P_i^O$  makes  $\partial c_{jr}$  dependent on  $i$  which is undue. This is solved by letting  $i$  varying over all the  $N$  neurons ( $N = 2$  in this specific case) in the output layer so that:

$$\frac{\partial E}{\partial y_j} \frac{\partial y_j}{\partial P_j^H} \frac{\partial P_j^H}{\partial c_{jr}} = \sum_{i=1}^N \left( \frac{\partial E}{\partial P_i^O} \frac{\partial P_i^O}{\partial y_j} \right) \frac{\partial y_j}{\partial P_j^H} \frac{\partial P_j^H}{\partial c_{jr}}$$

And considering that  $y_j$  is an input to neuron  $i$  in the output layer, it is possible to chain  $\partial u_i$  (differential of the output signal of the neuron  $i$  in the output layer), obtaining:

$$\sum_{i=1}^N \left( \frac{\partial E}{\partial u_i} \frac{\partial u_i}{\partial P_i^O} \frac{\partial P_i^O}{\partial y_j} \right) \frac{\partial y_j}{\partial P_j^H} \frac{\partial P_j^H}{\partial c_{jr}}$$

Which can be simplified in:

$$\sum_{i=1}^N ((t_i - u_i) f'(P_i^O) w_{ij}) \frac{\partial y_j}{\partial P_j^H} \frac{\partial P_j^H}{\partial c_{jr}} = \sum_{i=1}^N ((t_i - u_i) f'(P_i^O) w_{ij}) f'(P_j^H) x_r$$

Finally, the training performed for all  $R$  patterns, leads to define the increment  $\Delta c_{jr}$  as:

$$\begin{aligned} \Delta c_{jr} &= \eta \sum_{k=1}^R (\bar{y}_j^{(k)} - y_j^{(k)}) f'(P_j^{(k)}) x_r^{(k)} \\ &= \eta \sum_{k=1}^R \left( \sum_{i=1}^N ((t_i - u_i) f'(P_i^O) w_{ij}) f'(P_j^H) x_r \right) \end{aligned}$$

It is possible to define the process for any number of hidden layers. Let us define a perceptron with  $L$  hidden layers, each one composed by  $M_l$  neurons, and an output layer with  $N$  neurons.  $w_{ij}^{(l)}$  is the generic weight of  $i$  neuron in layer  $l$  with respect to the neuron  $j$  in the layer  $l - 1$ . So  $\Delta w_{ij}^{(l)}$  is:

$$\Delta w_{ij}^{(l)} = \eta \sum_{k=1}^R \delta_i^{(l),(k)} e_j^{(l-1),(k)}$$

Where  $e_j^{(l-1),(k)}$  is the output signal of the  $j$  neuron of the  $(l - 1)$  layer in correspondence of the  $k$  pattern of the training set with:

$$\begin{aligned} \delta_i^{(l),(k)} &= (t_i^{(k)} - u_i^{(k)}) f'(P_i^{(k)}) \quad \text{if } l = L \\ \delta_i^{(l),(k)} &= f'(P_i^{(k)}) \sum_{r=1}^{M_{l+1}} (\delta_r^{(l+1),(k)} w_{ri}^{(l+1)}) \quad \text{if } l < L \end{aligned}$$

$M_{l+1}$  is the number of neurons in the  $(l + 1)$  layer. The overall procedure for a single iteration step can be then synthesized as follows:

1. Start from  $l = L$ , so the output layer, and compute  $\delta_i^{(l),(k)}$  for each neuron  $i$
2. Move to  $l = L - 1$  and compute  $\delta_i^{(l),(k)}$  in function of  $\delta_i^{(l+1),(k)}$
3. Continue to compute  $\delta_i^{(l),(k)}$  until  $l = 2$
4. Update all the weights  $\Delta w_{ij}^{(l)} = \eta \sum_{k=1}^R \delta_i^{(l),(k)} e_j^{(l-1),(k)}$

Finally, a method to train a multilayer network in a supervised learning paradigm is obtained.

The supervised learning method just proposed is a good way to train an ANN to perform tasks that are similar to the ones presented in the 2.1 *Machine*

**Learning** chapter, like classifying patterns, map variables into homogeneous groups or approximate functions. So, it would not be wrong to say that the ANN could be considered another technique available to carry out such tasks, making almost excessive the attention given so far to the description of this learning approach. In addition, following the learning rules given for ANN, someone could argue that the same problems of ML could arise again: indeed, also in the supervised training of an ANN, the choice of the training set has a strong influence on the final result of the learning procedure, so the human intervention is fundamental to reach good performances and complicated tasks, like image recognition, are still not manageable.

The same consideration can also be made in case you move into the unsupervised learning field: this approach is available also in the ANN, but the process leading to the construction of a network capable of solving the typical unsupervised problems (clustering, anomaly detection, and so on) is longer and more complicated than the supervised one. For instance, determining the number of neurons in the hidden layer is an extremely significant issue in unsupervised neural network design, and selection of hidden neurons randomly may cause the problem of either Under fitting or Overfitting in a network [21]. For this reason, only the supervised learning approach was presented.

As we said earlier, the ability of an ANN to specialize in a specific task is defined by the combination of weights and threshold, but also by the structure of the network itself: indeed, specific topologies offer some different solution to overcome the problems indicated before, so depending on the task required, it may be better to apply a specific architecture than another. A huge number of structures exists, so it is impossible to describe the development of each architecture, and the description of every single method is outside of the scope of this thesis. The most diffused architectures of such ANNs are the convolutional and the autoencoder neural networks: they both are feedforward networks aiming at learning a compressed,

distributed representation (encoding) of an input dataset (usually an image but in principle can be applied to any generic input pattern) [20].

In the following, the convolutional networks will be described in detail, as example of an ANN widely used in different field. This structure finds a lot of applications, due to its ability to analyse images and extract from them information in a fully automatic way: also, in medical application the Convolutional Neural Network (CNN from now on) are used in all those medical branches which exploit imaging techniques.

### 2.2.3 The convolutional neural network

When the ANNs was firstly described in the earlier pages, it was said that this technique aims to reproduce the human behaviour. The vision system in our brain works in a complex way, in which different parts of its structure work in series to understand the image arriving from the retina of our eyes. From the primary visual cortex V1, up to the entire series of virtual cortices from V2 to V5, each part implements a more complex image processing [20].

Indeed, from the image captured by our eyes, the brain carries out a series of compressions that have the purpose of extrapolating from the image the fundamental characteristics that allow to identify the objects present in the scene, the distance between the objects and the body, the background and so on. So, the analysis goal is to pass from the “specific object in the scene”, for instance “the red, small apple on the desk”, to “a more abstract concept” collected in our memory because the brain had already faced that thing, for instance “an apple”.

The CNN is an ANN topology that tries to mimic the brain structure, and its way to process the image. This neural network is basically composed by three types of layers: a **convolutional layer** followed by a **pooling layer**, and finally a **feed forward fully connected** one [22]. The first one is responsible to the extraction of the main features of the input images and learns the characteristics of the presented scene; the latter is a classical classification network that is able to define which things are present in the input image, exploiting the features offered by the previous stage.

The convolutional stage follows the connectivity pattern of neurons in our brain, and in particular of the visual cortex: individual neurons respond to stimuli only in a restricted region of the visual field, known as *receptive field*. These fields slightly overlap, to cover the entire visual area, but sharing a little bit of each one to increase the accuracy of the analysis [23]. Starting from this structure, the CNN basic ideas are *local receptive fields*, *shared weights*, and *pooling*: the first two are exactly the ideas exposed earlier, so the receptive fields and their overlapping, while the pooling is a process of compression of the image, done to reduce the computational power required by the brain, but also by our artificial systems.

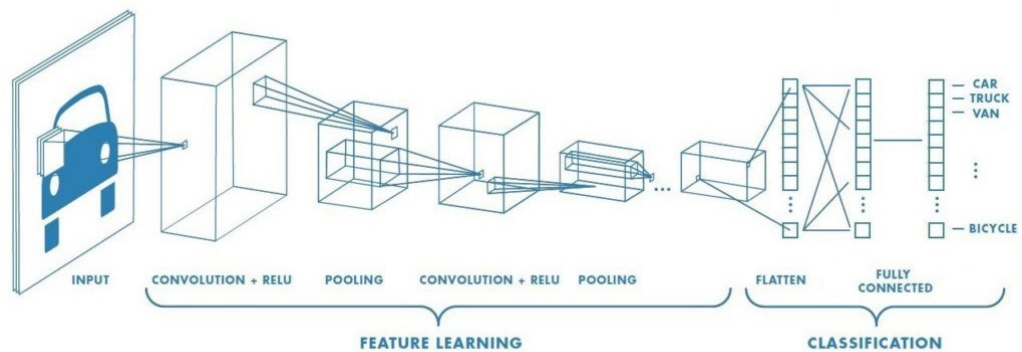


Figure 18: example of CNN

The behaviour of the example CNN in Figure 18 can be divided in four parts:

1. The input layer, as in the common ANN, collect the image pixels, that are the input values of the CNN.
2. The convolutional layer will determine the output of the neurons belonging to the different local receptive fields, through the product between the weights and the region connected to their input. The **Rectified Linear Unit (ReLU)** is a stage in which an activation function tries to force the output of the layer to be positive, with the aim of normalizing the values of the convolutional layer output.
3. The pooling layer will then perform a downsampling of the convolutional layer, to reduce the dimensionality and lower the computation required.
4. The fully connected network at the end performs the usual tasks of an ANN.

So, the process made by a CNN starts with the convolutional layer, characterized by the use of the local receptive fields. This is a set of connections that share the same weights and thresholds and are replicated on the entire visual field forming a **feature map**. Each local receptive field is computed by a neuron, and each one take as input a specific number of inputs from the previous layer, like depicted in Figure 19.

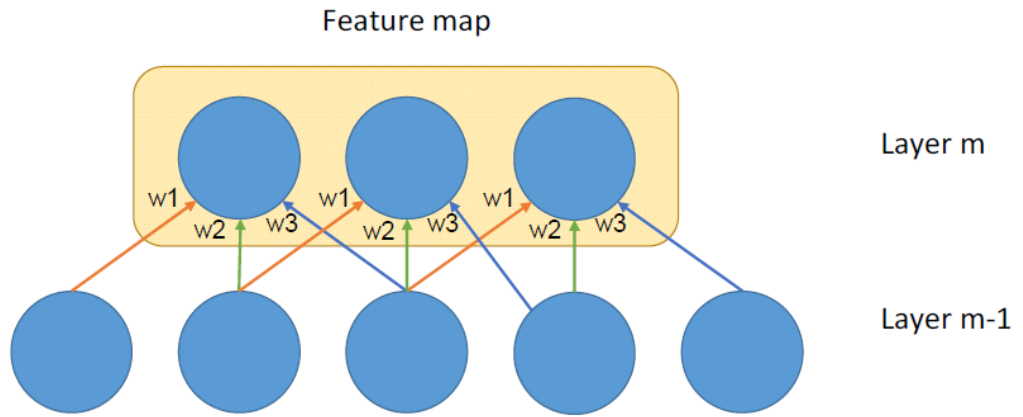


Figure 19: example of a feature map

In other words, a feature map is composed by a set of neurons, in which each unit analyses a specific part of the image, through the application of a filter called **kernel**. Each neuron computes a new pixel, function of the previous ones, and the collection of all the outputs generate the feature map. The process to obtain this map is the convolution of the input image with a linear filter, the summation of a bias (the threshold, considered as a virtual input of weight equal to  $-1$ ) and finally applying the activation function.

$$u_{j,k}^{(i)} = \tanh \left( b^{(i)} + \sum_{l=1}^{f_x} \sum_{m=1}^{f_y} w_{l,m}^{(i)} a_{j+l,k+m} \right)$$

Where  $f_x$  and  $f_y$  are the dimensions of the kernel, so the dimensions of the pixel matrix analysed by the filter,  $u_{j,k}^{(i)}$  the value of the feature map at layer  $i$  and in position  $j, k$ ,  $b$  is the threshold,  $w$  is the matrix of the weights that performs the filtering,  $a$  the value of the pixels of the local receptive field.

So, the kernel is moving over the image and the scalar product between its value and the pixels one is computed, then the bias is summed up, and the new pixel value is calculated through the activation function. If, for instance, a 5x5 pixel image passes through a convolutional layer with local receptive fields of dimensions 3x3, the first neuron would provide:

$$u_{1,1}^{(i)} = \tanh \left( b^{(i)} + \sum_{l=1}^3 \sum_{m=1}^3 w_{l,m}^{(i)} a_{1+l,1+m} \right)$$

For a total of  $3 \times 3 + 1 = 10$  weights to compute for each neuron (nine weights and one threshold).

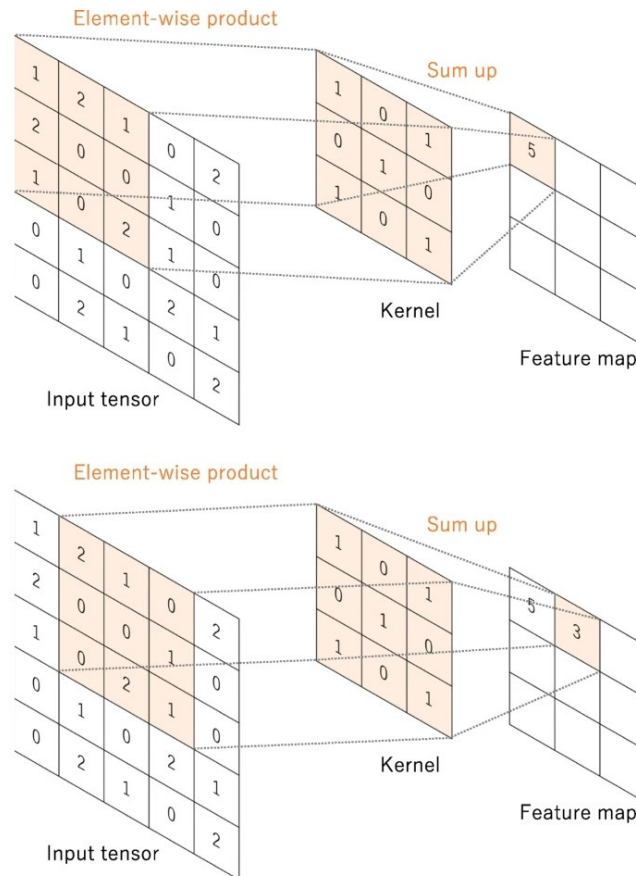


Figure 20: example of a kernel moving on an image



As depicted in Figure 20, assuming that the kernel moves from left to right and from up to bottom, one pixel at a time, it would be necessary to take three steps to reach the upper right edge, then again three steps to reach the bottom right corner, for a total of  $3 \times 3 = 9$  receptive fields, so nine neurons for the layer. The number of parameters to compute at the end of just one layer, and for a small image like the one in the example, is  $10 \times 9 = 90$  parameters to define. As image resolutions are usually higher than this one and as multiple layers are usually applied in CNN, it can be already noticed why the computational power required for this kind of application is so huge.

An example of the result of this process is shown in Figure 21, where different feature maps are presented, following the downsampling process: indeed, it is possible to notice that the dimension of the pixels increase in the process, so the details on which the feature map is focused on become more and more discernible. In the upper sequence the contours of the hair are highlighted, while in the bottom one the eyes of the woman are enhanced.

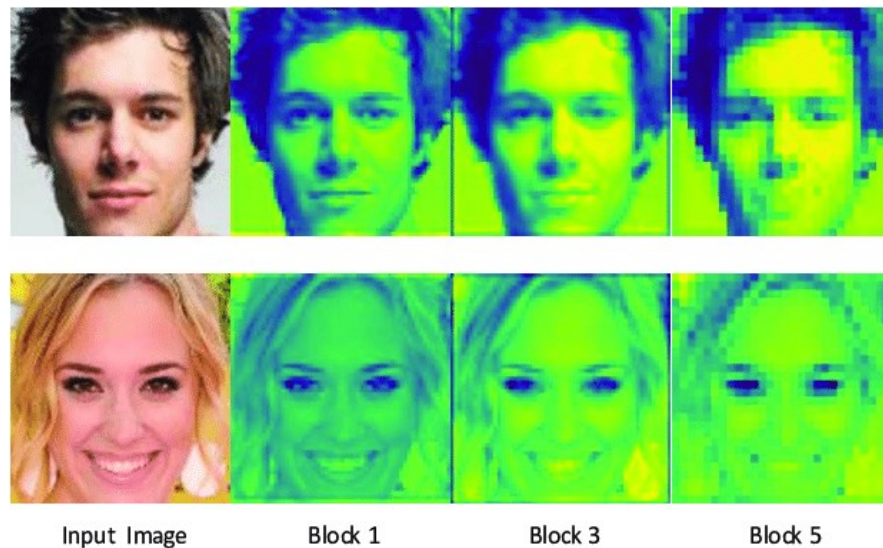


Figure 21: example of different feature maps

After this process the dimension of the image is reduced, as the feature maps generate one pixel each step, so we move from a  $5 \times 5$  image, to a  $3 \times 3$  one, that enhanced one

specific feature of the image: the feature map can highlight the contours of the main object in the image field, or maybe some details in the scene.

The next passage is the ReLu, which is necessary in order to obtain from the convolutional layer only positive values: indeed, the activation function inside each neuron can generate also negative values, for instance it was previously used the *tanh* activation function that generates value in the range  $[-1; +1]$ . The rectified linear unit is a function defined as:

$$f(x) = \begin{cases} x & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

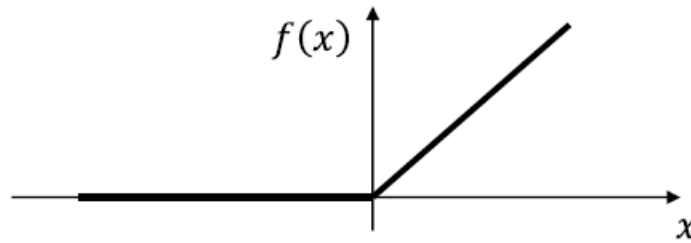


Figure 22: example of ReLu function

After this process, all the output values of the ReLu layer, which will become inputs for the next layer, will be all positive.

Finally, the last process is the pooling, so the reduction of dimensionality in order to lighten the computational load, but also to introduce a translation invariance to small shift and distortion [24], so the feature extracted are not dependent on their exact position and are less sensitive to noise. It is a non-linear downsampling and there are not learnable parameter or process, as all the pooling characteristics are fixed. The pooling techniques are similar to the convolutional layer process, so a kernel of dimension  $M \times M$  flows on the matrix and for each step it reduces the number of pixels to just one, and how the kernel select the output values depends on which pooling method is chosen. There are two types of this technique, called **max pooling** and **global average pooling**: in the first one, shown in *Figure 23*, the kernel output is the highest value inside its field of view, while in the latter is an extreme

downsampling, in which a feature map is downsampled into a 1x1 array by simply taking the average of all elements in each feature maps. This is usually applied only once before the fully connected layers, to reduce the number of parameters to be learnt and to enable the CNN to accept inputs of variable size [24].

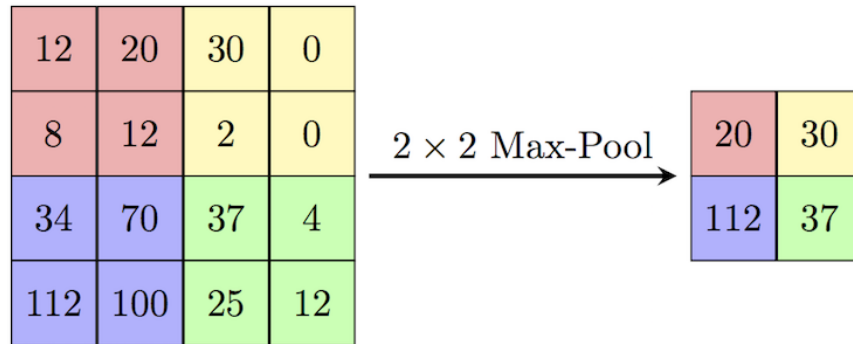


Figure 23: example of application of max pooling

As shown in Figure 18, several layers are present in a CNN, each one made by the combination of convolutional layer, ReLu and pooling. The basic idea is to chain different blocks with a decreasing number of units, up to a final stage of feature classification, which usually is a fully connected feed forward network able to assign classes depending on the set of features provided by the convolutional stage.

The reduction of the number of units, proceeding through the hidden layers, aims to increase the generality of the features extracted by the different layers: the first neurons in the network analyse small details, like contours and eyes like in Figure 21, and progressively the following layers try to merge these small characteristics, trying to compose larger and larger structures. These are then exploited to recognize the different object present in the image, and finally exploit them to identify the subjects of interest in the image field.

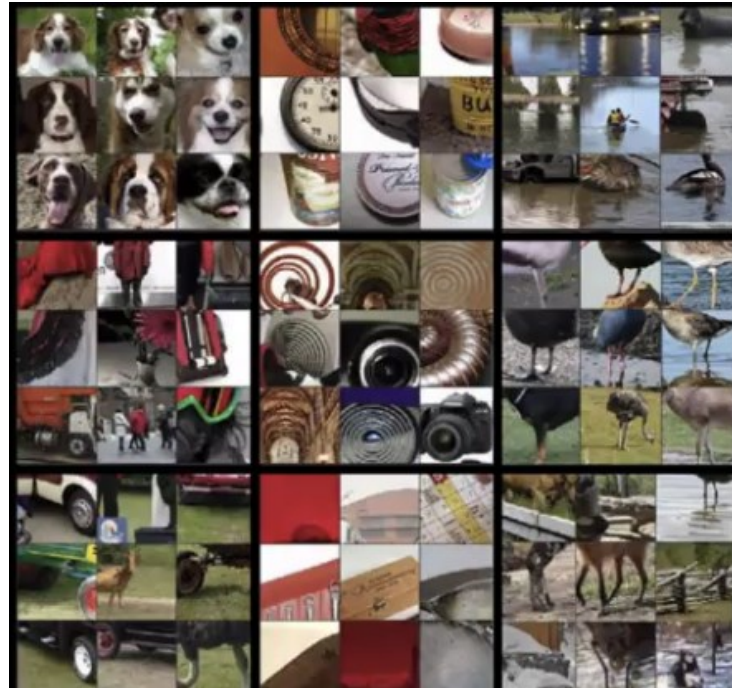


Figure 24: example of what the deeper layers of a CNN learn

In Figure 24 are shown the different structures learnt by a CNN in the deeper layers, which learnt to recognize some dogs or also the wheels of a car, and it can use all these knowledges for different purpose: indeed, the final result of the CNN depends on the final network in series to the CNN. For instance, the features extracted from the network can be exploited to classify some images, or to generate new ones mixing the features learnt in the training, or also to separate a set of images depending on some important features, always discovered by the system independently.

## 2.3 AI in medicine and healthcare

Artificial intelligence is now exploited in almost every area in the world, of course also including medicine and healthcare: assistance to physicians in diagnosis and treatments application, analysis of public health systems or the study of data in clinical trials.

The approach used in medicine is based on the knowledge and experience of clinicians, which generate usually qualitative assessments: these evaluations suffer

from problems related to the human sphere such as fatigue, misunderstanding of links between symptoms or the inability to notice very small details. In contrast to such qualitative reasoning, AI excels at recognizing complex patterns in imaging data and can provide a quantitative assessment in an automated fashion [25].

In particular, AI are really helpful in those field in which is required to interpret large amounts of data, and from them extract high level correlations difficult to understand by a human. For instance, to assist the physicians in the decision process that lead to diagnosis or treatment plans: usually is necessary to face these problems in a multi-parametric approach, in which different kind of exams are done, and the information extracted are combined to obtain a single result. Exams can be imaging technique, chemical analysis or notes taken by an expert from a qualitative point of view, and an extreme workload pours on the shoulder of doctors which must take the final decision. AI can be really useful in this application, exploiting the speed with which these machines are able to extract the most significant information from a massive dataset, or even the ability to analyse biomedical images and report the presence of anomalies, and so on.

Using ML or DL approaches can improve the general level of the clinical practise, making it more efficient, personalized, and more convenient. Indeed, it is now possible to produce biomedical instrumentation able to monitor the single patient in a more customizable way than in the past, and so the amount of information available for each patient is grown. People is beginning to demand faster and personalized care [26] [27], due to their perception of a higher technological level, that creates the idea of a totally automated process that work without a doctor. This is not true, but it necessary to help physicians to follow the patients request in faster diagnosis and solution, without any loss in efficiency and accuracy.

In cardiology, the ML technique is used massively to discover relationships between the independent variables (the information on the patient) and the dependent one (presence of a pathology) [28]. Supervised learning methods are used for different purpose: for instance, in [29] Halim et al. try to predict myocardial infraction or deaths, exploiting the combination of medical variable and proteomic measurements in a regression algorithm. Another example is the work made by Cui et al. [30] in

which they evaluate the diagnostic utility plasma biomarkers for in-stent restenosis (ISR), using a Support Vector Machine algorithm that reach an accuracy of 90%.

The DL approach, on the other hand, is widely used in the field of Image-based diagnosis, due to the efficiency of these algorithm in the analysis of images: radiology, ophthalmology or dermatology are just examples of the medical branches that exploits imaging techniques.

In radiology, usually diagnoses are made after acquiring different types of images, like radiography, MRI, PET, and so on, and the radiologists use these images for screening and making diagnoses or tracking the patient respond to a treatment [31]. In this area, the image analysis ability of deep neural network can be used to assist clinicians in their job: for instance, the first FDA approval for a machine that exploits deep neural network [32] was the Arterys [33], a system for diagnosing cardiovascular diseases which uses cardiac MRI images analysed by a deep NN trained with thousands of cases, and it continues to improve itself each time it is used.

Dermatologists can use AI in the skin melanoma diagnosis: in the classical approach, a rule of thumb is usually used, and it is called ABCDE, with each letter representing one criterion that influence the result: these criterions, with the exception of the E one, can be implemented extracted from a single image of the lesion under analysis, but in the last years different automated system were developed, like the Esteva et al. [34] one, which use a CNN trained with more than 100 thousands images, and achieved the same accuracy level of a test dermatologists group.

Other applications of AI in medicine are available, but the presentation of all these methods is out of the scope of this thesis. The above pages are indicative of the level of diffusion of AI within the medical sector, but despite the results that are obtained are substantial, there are some problems which limit the use of these technologies: some of these problems are of general term, such as privacy related to the data collected, while others are specific to the medical application, such as legal responsibility in case an automated system fails in diagnosis.

Another important problem is the way these AI methods work most of them can be represented as black boxes, so machines that take inputs and generate outputs, without showing the process that led to the result. While in some application this

could not be a problem, in the medical area there are requirements which cannot be ignored: indeed, the information exploited in medical applications is sensitive data, which contains very personal patient information. From a legal point of view, not only is it necessary to know the way in which these techniques manipulate data, but it becomes even more important to know whether these retain sensitive information in some way.

It is also important, from a practical point of view, that the patient is aware of what examinations he makes, what result can be expected and that, in general, the patient is confident of the treatment that the doctor builds on the information available to him.

All these problems will be more studied in the next chapter, in which will be presented the solution that is currently spreading as state of the art in the field of artificial intelligence, when the requirements shown above are mandatory, like for example in medical applications. The solution is called Explainable Artificial Intelligence (XAI), that try to solve the problem of interpret how the artificial intelligence internally works and the output it generates.

### ***3. Explainable Artificial Intelligence***

The concept of Artificial Intelligence, as already said in this thesis, has the aim of create a model of our intelligence, trying to replicate its behaviour. One of the most powerful aspect of human thinking is the ability to **learn from experience**, exploiting the daily life situations in a constant attempt at improvement. Another fundamental ability is the possibility of people to **explain** what they decide to do: in a world governed by causality, justifying one's own actions is one of the basic concepts of interpersonal relations. In fact, each of us has completely different behavioural patterns from the others, so it is often necessary to point out the motivations that have pushed us in one direction, to allow others to understand the reason of our choices. While the learning concept, as it had been showed, is already a standard in AI, the **explainability** or **interpretability** of the methods is still a difficult result to reach.

However, Artificial intelligence, as said in the previous chapter, is rapidly spreading around the world, finding applications in almost every field possible, like home assistant, self-driving car, healthcare. The attention of this thesis will from now focus on this last category: in medicine, artificial intelligence helped to upgrade the so called *medical technology*, so a “range of tools that can enable health professionals to provide patients and society with a better quality of life by performing early diagnosis, reducing complications, optimizing treatment and/or providing less invasive options, and reducing the length of hospitalization” [35]. AI moved the fundamentals of the medical technology from the classical instruments like prosthesis, stents and so on, to a more “mobile” era, helped by the diffusion of mobile and wearable devices, internet connectivity and smartphones, that augmented the capability of implementation of artificial intelligence applications in a more easy and reliable way.

This automation process is still in its infancy, as the real applications of artificial intelligence in the routine clinical life are not so many, and typically their tasks are to assist human physician in making diagnosis or planning treatments, or in the wearable applications where devices are able to collect some health information through sensors and provide visualization of data and suggest “healthier” lifestyle.



Always in the above pages, it was said that in particular field like the medicine one, the application of AI technologies requires particular attention, due to the delicate aspect that these tools need to fight: accuracy in their predictions to avoid misclassifications of dangerous situations for patients (tumour, lesions or pathologies in general), clear and unique results given in each situation and so on.

Expanding our point of view, we will consider both the already applied algorithms and the ones under research. These technologies still struggle to establish themselves as standards in clinical practice, as they present several problems that restrict the application of these tools in an easy and safe way: an accuracy that in many areas is not yet able to exceed that of an experienced doctor, legal uncertainties in the event of adverse events, a lack of confidence on the part of staff who must necessarily update their knowledge to cope with a completely new category of tools.

However, the main problem is to be found in a poor ability of the developers of these systems to consciously handle all aspects of these automated technologies. In fact, as things actually stand, it is possible to find a correlation between the ability of the AI technology to perform complicated and delicate tasks and their complexity: an example of this characteristic is given by the usual behaviour of an artificial neural network, that becomes able to improve its data imaging capabilities as it increases its hidden levels and neurons. This fact was confirmed by the discovery of the *double descent phenomenon*, in which the accuracy of an algorithm of machine learning (but even more for deep learning) follows a sort of typical behaviour: as the number of parameters of the model, or the number of epochs in the learning procedure or the dimension of the dataset is growing, the accuracy firstly increases, then it decreases and finally starts again to augment.

This phenomenon was discovered in the last decades, since the machines calculation speed was not high enough to get the second increasing climb within a reasonable time, so the algorithms typically were stopped at the first sign of inaccuracy. As the computational power increased, speeding up the calculation, machines became able to compute models with a higher number of parameter or epochs in a faster way, allowing to overcome the decreasing zone and showing the second accuracy gain. Indeed, almost every method has an initial increase of the accuracy, then it gets worse

and after a threshold it start to increase again as the model fits the data in a better way. This phenomenon is depicted in Figure 25.

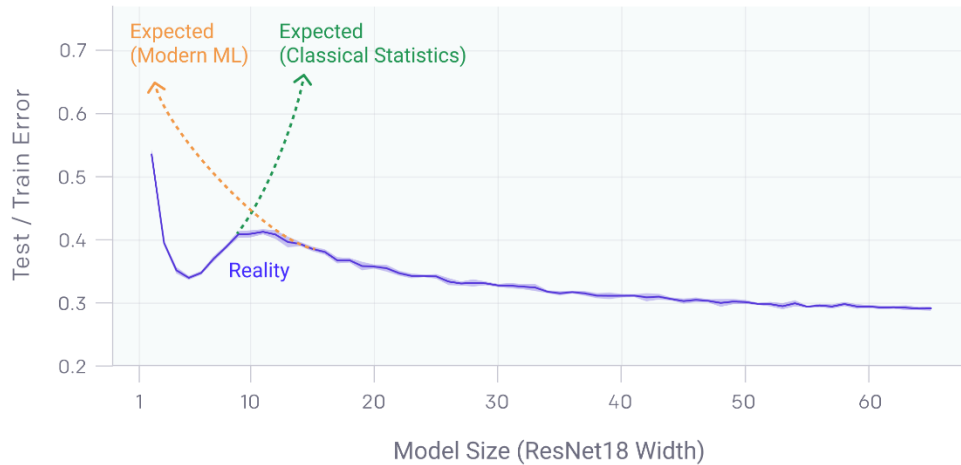


Figure 25: double descent phenomenon

The green curve is the representation of the theoretical framework of the classical statistics, which states that increasing the number of parameters the model become continuously worse, while the yellow one is the expected behaviour of a modern ML model, that increase its accuracy augmenting the number of parameters as the algorithms starts to overfit on the training data. In the practical world, as the amount of data in not infinite, and the computational time requires a finished number of parameters, the ML algorithms need to be stopped at a certain point depending on the complexity of the system.

So, following this phenomenon, someone could say that if we want to solve complicated problems, algorithms must be complicated too: as this approach has been followed in the last decade, the most of the working AI algorithms are under the denomination of **black boxes**. This name comes from the impossibility to understand completely the way in which the AI method is working internally: a high number of parameters, a complicated system topology, and more in general “the way in which the program processes the data to produce the outputs” it is not fully *explainable*, also by the experts that made the algorithm itself.

The “black box problem” is the most difficult issue to solve, as this was taken into account in the recent years, given the current spread of AI in every aspect of our lives. In fact, as our lives are surrounded by increasingly intrusive AI technologies, it is necessary that these are comprehensible at least from the point of view of how they work. In particular, this need becomes mandatory in those contexts in which the choices made by machines influence situations potentially dangerous to people. In the medical context for example, it is unthinkable that algorithms which assist doctors in diagnosis work in a more or less opaque way.

This issue became so important that in 2018 the European Commission, for instance, introduced the General Data Protection Regulation (GDPR<sup>2</sup>), establishing that companies placing artificial intelligence instruments on the European market must be able to explain the choices made by these instruments.

Words like *explanation*, *interpretability* or *explainability* became in the recent years a trend, when talking about artificial intelligence. Documents like the GDPR, *highlighted* the need to retract the way we interface with the AI technologies with which we live: in fact, in contexts in which automated activities have a weight on decisive choices on people’s lives, like in the medicine field, you need to know how these machines calculate their results. For these reasons the actual trend, when dealing with AI technologies applied in delicate and dangerous field, is to put the attention on the interpretability/explainability of these methods, passing from the simple Artificial Intelligence to **Explainable Artificial Intelligence (XAI)**.

The Defense Advanced Research Projects Agency (DARPA<sup>3</sup>: a research agency of the United States Department of Defense) says [36] that the Explainable AI (XAI) program aims to create a suite of machine learning techniques that:

- Produce more explainable models, while maintaining a high level of learning performance (prediction accuracy).
- Enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.

---

<sup>2</sup> Document in italian available at <https://eur-lex.europa.eu/legal-content/IT/TXT/HTML/?uri=CELEX:32016R0679>

<sup>3</sup> [https://it.wikipedia.org/wiki/Defense\\_Advanced\\_Research\\_Projects\\_Agency](https://it.wikipedia.org/wiki/Defense_Advanced_Research_Projects_Agency)

This kind of artificial intelligence has the aim to create algorithms whose results are more understandable to humans [37]. there are several techniques that systems can use to explain their results: some explanations can be in the form of natural human language, then sentences logically linked to each other; or in visual form as diagrams or heat maps, that seek to give a direct view of the characteristics that are weighing most on the results provided by the system. A comparison between a normal AI and the XAI potential results is depicted in Figure 26.

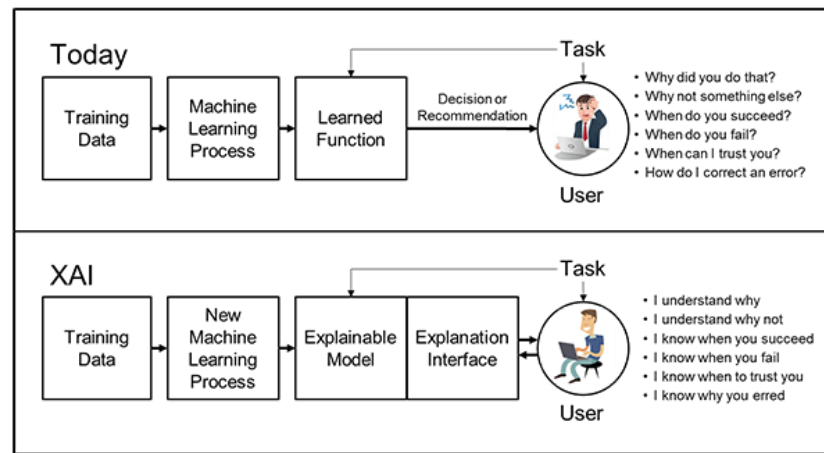


Figure 26: comparison between AI and XAI

The different explanations must also consider the target of the explanation. Understanding a concept is not an objective process: on the contrary, the way in which each person understands something is extremely subjective, and depends on their own culture, experience, and character. Moreover, within the same context, explanations of the same concept may be necessary, but referring to recipients all different from each other. Let's say the case of an automated system that wants to assist the doctor in deciding a pharmacological treatment on a patient, and you want to use an XAI that gives explanations to both the doctor and his patient: it is obvious that the explanations must have two different complexities, because the patient does not have the cultural background of the doctor; they must give different information, because the patient may be interested in more certain information that instead the doctor ignores and so on. So, the definitions of interpretability and explainability are, thus, domain dependent and may not be defined independently from a domain [38].

In literature usually interpretability and explainability are used as synonymous, while some authors like [39] make a separation between the two terms: in this thesis will be followed the first approach.

In the recent years the spreading of AI systems in decision-making processes raised the attention on this kind of technology, as illustrated in Figure 27: in this image is shown the Google Trend<sup>4</sup> search analysis of the keywords “Explainable Artificial Intelligence” in the last 5 years; each point represents the weekly rate of research for this word, and it is possible to see that is constantly growing. The explainability of a system is fundamental because it allows to explore the mechanism that the different algorithms exploit to generate their decisions.



Figure 27: google trend for "explainable artificial intelligence" of the last 5 years

In this way, experts can identify errors more easily and accurately, both in case a bug prevents the machine from working, and in the case where a wrong configuration leads to misclassifications. But the knowledge generated by an XAI also allows to make the algorithm more reliable to the end user, who understands better what the algorithm suggests to him, which is fundamental in some fields, particularly in medicine: a high level of confidence between doctor and patient is essential in the process of diagnosis and treatment of a disease, as the patient is determined to follow what the expert suggests to him to do; likewise, in the eyes of the patient, a medical assistance technology will be more reliable if it can explain how it made its decisions,

---

<sup>4</sup> <https://trends.google.it/trends/?geo=IT>

because it will not be only the output generated by a cold blood machine, but this result is followed by some explanations of how it decides.

In the next pages, we will try to understand which are the consequences of passing to this kind of technology, if there are issues and finally how these technologies are applied in the context of this thesis, namely the medical one.

### 3.1 Why use an XAI

The motivations that push for the use of an XAI compared to a classic AI are many and arise from several issues related to the use of these automated technologies: privacy problems due to the unknown way in which algorithms collect our personal data; or problems for insurance or legal liability issues of any adverse events related to the choice of machines, or even establish insurance policies related to the use of artificial intelligence tools. In general, it can be said that the use of XAI is essential in all those applications in which the end user needs to fully understand the results obtained by AI, so that they can be managed optimally.

Four main objectives are identified by [40] for an XAI: trustworthy, confidence, transparency and informativeness.

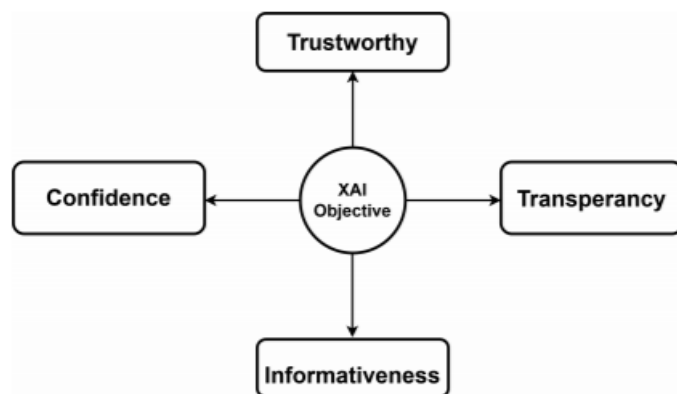


Figure 28: main objectives for an XAI

The concept of transparency is justified by the attempt of the XAI to make black box models less opaque, in the sense of making more accessible the mechanisms that exploits the algorithm to generate the outputs. Informativeness can be considered a direct consequence of the less opaque algorithm: models can be more consciously analysed, corrected, improved.

Trust is based on the fact that in our lives we tend to rely on things we know how to explain how they work, so is the XAI.

The confidence arises from the combination of the previous objectives: if we are in a context we know, we are led to feel more confident about ourselves. In the same way, the use of a technology we can trust, of which we know the characteristics, and which has few dark sides, allows us to be more confident in its use.

In [37] a more technical analysis is done, but again at least four reasons are found to justify the use of the XAI instead of classic AI: *justify, control, improve, discovery*.

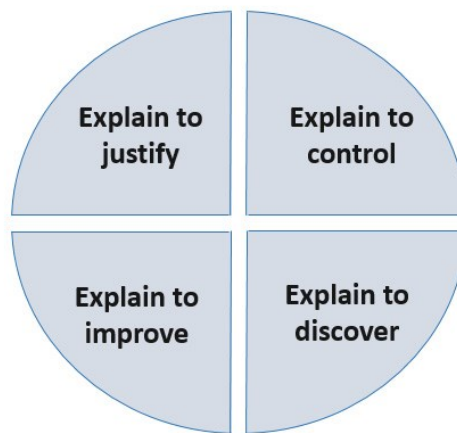


Figure 29: reasons to use an XAI

### 1. *Justify*

In our lives, justifying our actions allows us to relate to other individuals, trying to understand the reasons that lead one person to make a decision with respect to another. It is not a matter of understanding the whole decision-making process, but simply which were the basic conditions that led to make a certain choice, even more if these alternatives influence particular aspects of the lives of the people around us.

For these reasons, it is no longer acceptable that automated systems involved in decision-making processes related to sensitive issues, such as the diagnosis of a disease, are completely black boxes. It is necessary that AIs are able to tell which were the main conditions that led to generate a certain output: In this way also the confidence that develops around a given result can be evaluated on the basis of the motivations that have led to it.

For instance, imagine the situation where a mobile device must report the presence of a dangerous situation. The reliability of this result is certainly greater if the device accompanies its decision with indications on the reasons that led to this decision, compared to a situation in which the system simply generates the alarm

## *2. Control*

Knowing which are the motivations behind a choice is essential also in the evaluation of that decision: if we explain our actions before we take them, it is easier for us and others to give an opinion on what we are about to do. This allows us to have more control over our actions and those of others, allowing us to prevent any mistakes or encourage correct attitudes

The same requirement must be asked to AI technologies involved into decision making processes: knowing how these machines did their choice allows us to detect and correct any errors, which in practical application could bring extremely dangerous conditions.

Let's get into a medical context. An automated system, composed by a complicated convolutional neural network, is being planned to diagnose an oncological pathology through the analysis of tomographic images: the system starts to work effectively in the training dataset, but in the test phase accuracy precipitates in an unacceptable manner. If the system is a simple AI, it becomes extremely complicated to understand how the system makes its decisions, and consequently correcting the algorithm becomes an unsustainable job. On the contrary, if the system is an XAI, we have the possibility to understand which are the causes of errors through the analysis of the reasons behind the choices made by the system.

## *3. Improve*



With the same premises as the control problem, the explanations provided by an XAI allow to analyse the motivations behind his choices with the aim of improving his organization in a more intelligent and efficient way: for example, you can understand that certain features have a decisive impact on the final decision, so you can think of lightening the system by ignoring the already ignored basic features, or combine basic features into a more generalized one by reducing the number of model parameters and so on.

#### 4. *Discover*

In the *Artificial Intelligence* chapter, it has been said that ML models are not really programmed, but they are allowed to learn as independently as possible. For this reason, it is not said that the knowledge generated by these systems faithfully follows our way of interpreting reality. This situation can be extremely favourable in the context of an XAI that explains how it has reached its level of knowledge, which may perhaps allow us humans to discover things that we had not previously noticed.

For example, we place ourselves in the situation of a clinical study that tries to correlate the clinical history of a patient with the possibility of being affected by a particular pathology: once the system has reached a high accuracy, it is possible to study the motivations behind its output, and maybe discover a link between a particular personal characteristic (whether physiological, pathological or lifestyle) and pathology. This situation is possible only if using an XAI, that explain how it reached its results.

The four reasons given above may be misleading: in fact, one might be led to think that it is always advisable to use an XAI, as they add to the already high capacities of AI the possibility of obtaining justifications and thus having the tools to further improve their abilities. This conviction, however, immediately clashes with an inconvenience that appears in the practical world.

Indeed, the main problem of XAI systems is that they present a sort of trade-off between accuracy and interpretability, that must be considered. There is not a theoretical formulation that states that this trade-off exists, but the actual state of the

art in many domains indicates that in the practical world, this trade off exists and must be taken into account.

In the previous chapter it was said that the ability of an AI algorithm to perform complicated tasks, is somehow proportional to its complexity, so for very hard problems are required very complicated algorithms. If the aim of XAI is to give an explanation of how these methods are working, it is obvious that it is impossible to have an intricate algorithm and a sufficient interpretation of its work at the same time: imaging an artificial neural network, for example a CNN for computer vision, in which the complexity in terms of its topology is necessary to reach a better data analysis and feature extrapolation, so it is very difficult to explain how each hidden layer is doing in the normal working; to gain more interpretability we should reduce the number of layers, or the number of neuron of each level, so maybe we could better understand the network job, but effectively losing accuracy, as the structure of the CNN influence a lot the algorithm performances.

For this reason, the XAI is not a fundamental tool, which must be always implemented, but it is more a strategic resource to be used in all those contexts in which it is necessary to understand in depth the tools that are being used. Its application depends mainly on how opaque the system is, how dependent the context is on the automated system, but above all on how the results influence very delicate decisions and how dangerous a possible error is. It is obvious that if a market study is being carried out on consumer tastes, there is no point in spending time and money on an explainable AI; In the same way, however, its use becomes essential when an automated system has to make decisions to support a doctor on the life expectancy of a patient for whom it is necessary to decide whether to carry out a delicate operation or not.

Now that the reasons why it would be useful to use XAI have been indicated, we can analyse the techniques currently used to implement these systems, in particular by referring to the way in which the explanations of the results can be generated.

## **3.2 How to implement explainability**

The variety of algorithms available within the ML is enormous, therefore also the ways in which you can express the explanations of a model are also numerous. Even in everyday life, we're used to adapt the way we explain things to the contexts that we're faced with, so based on the people we interface with, the context we're in, and so on.

Following the current literature and the state of the art it is possible to classify the different methods to generate explanations according to three principles [37]:

1. The complexity of interpretability
2. The scope of interpretability
3. Specificity on a particular model

These criteria are purely theoretical and try to give some order to a literature in continuous updating, with very high rhythms, which often make the understanding of methods chaotic and little intuitive, leading to not knowing well which method is best to choose for your application.

### 3.2.1 Explainability depending on complexity

The first way to distinguish the different methods is to group the algorithms according to the complexity of the model to be explained: as has already been said above, XAI fight a constant battle between the ability to be accurate AI and the ability to be interpretable.

Therefore, the most direct and simple option is to generate an algorithm *intrinsically interpretable*, so a model as simple as possible that can be directly interpretable but paying for this feature with a lowering of the overall accuracy. An example of this approach is analysed well in [41], where the concept of *Self-explaining AI* is proposed, as an AI model that provides two outputs, namely the prediction and explanation. In [42] are indicated two types of this interpretability, called *pure transparent* and *hybrid*: the first sub-approach we must use models that considered transparent intrinsically, like the work [43]; in hybrid family there is a combination of a transparent and black box models, trying to find a trade-off between

interpretability and performances in prediction, like in [44] in which some parts of a transparent model were substituted by small black box algorithms to boost the general accuracy.

The second option available is the so called *post-hoc explanation*: in this approach it is firstly generated a black-box model which is accurate, robust, and of course very complicated, then a set of different techniques are applied on the outputs of this AI, in order to give explanation to these results. It is like we are creating a second model, built on understanding how the first black-box algorithm produced the outcomes, in a reverse process in which we go from the outputs to the inputs, trying to understand which pathway was followed to generate the outputs.

Inside this category the explanation can be done in Natural Language, like the work done in [45], visualization of the model like in [46] and explanations by example like in [47].

The choice between the two approaches depends on how complicated the starting algorithm is: if a low number of parameters are used it is possible to apply the first method, but if the complexity increase it is better to use the post-hoc interpretability.

### 3.2.2 Explainability depending on scoop

When analysing automated systems such as those under observation, there are two points of view from which we can analyse their interpretability: the first is that of the general understanding of the functioning of the system, that is to understand the mechanisms internal to the model, that is the path that is followed in order to go from the inputs to each single output available; the second instead is that one to analyse every single output coming from the system to understand how this particular one was generated.

So, it is possible to distinguish two subgroups that follows these two approaches: the *global interpretability* and the *local interpretability*.

The first one refers to those models that try to make explainable all the logic behind their decision processes: however, these systems present problems, as they tend to

be strongly linked to their scope, becoming less general and therefore explainable only for their specific application. Moreover, we must always take into account the trade-off between accuracy and interpretability, so this type of models is really reliable only for a few internal parameters.

More promising is the group of methods belonging to the *local interpretability* group, in which we try to find explanations to the single output of the system, rather than the general logic of the model. Typically, the features that are highlighted are those that have contributed most to the definition of the specific output. This approach is very similar to our cognitive process, in which we tend to prefer the understanding of separate compartments and then tie them to form a general knowledge, rather than directly learning the whole process in a single solution.

Examples of this kind of approach are the local interpretability techniques used in the image classification task, in which the details extracted from the image and determinant for generating the output are shown through saliency maps, sensitivity maps or the pixel attribution maps, like the ones depicted in Figure 30.

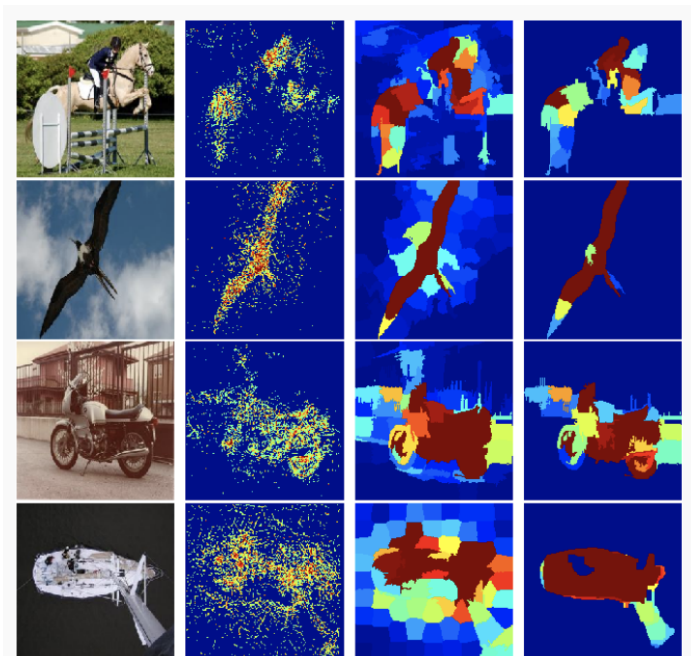


Figure 30: saliency maps in CNN image analysis

An interesting alternative is the possibility of mixing the positive characteristics of the two groups, to form one with common characteristics, but it is still under research.

### 3.2.3 Explainability depending on model dependency

In this last category, models are divided in *model-specific or agnostic*: indeed, to the first group belong all those models applicable only to a single type of ML algorithm, while in the second group those valid for any kind of algorithm.

The model-specific interpretability methods are strongly linked to the type of algorithm for which they were created. They have a robust correlation between result and explanation, so they are not very generalizable.

The agnostic ones are not tied to a particular type of ML, so they are highly generalizable and separate the prediction process to the explanation. To this kind of approach belong usually post-hoc methods, they could be local or general interpretable, and usually are used to explain ANNs. Being this category very general, the way these methods are implemented could be similar to the other ones presented before. There are four techniques available: *visualization*, *knowledge extraction*, *influence methods* and *example-based explanations*.

#### 1. *Visualization*

This method tries to illustrate the representation made by the algorithms, usually applied to deep neural networks in the supervised learning approach for images. There are different kind of techniques available, like the *surrogate models* that are interpretable model applied to black-boxes one and trained on the predictions of the latter one to understand the results; the *partial dependence plot* tries to graphically show the correlations between one or more inputs variables and the outcomes of a black-box algorithm.

#### 2. *Knowledge extraction*

This technique is mainly applied for ANNs and tries to extract some patterns and information during the training phase of a neural network: usually in this phase the ANNs learn several representations of the input data, so the idea is to extract these

patterns while they are learnt and present them to the end user. The available techniques are the *rule extraction*, that works in the training phase creating a sort of approximation of the input/output relation that the network generate, at the level of the individual unit, making the learning process more transparent; the other technique is the *model distillation* which make a compression of the model generated by the ANN in order to make it more interpretable reducing the number of parameters.

### 3. Influence methods

This approach tries to estimate the weight of a feature on the final result, changing the value of one of these features and computing how much the outcome changes in response to this variation. This kind of models are always visualizable.

The techniques are *sensitivity analysis* the measures how much the outputs are influenced by weights and inputs, varying their values, so working at the unit level; *layer-wise relevance propagation* backpropagate the output into the network computing how much the network maintain its parameter unchanged; finally *feature importance* calculate the variation of the final prediction error in response of parameters variation.

### 4. Example-based explanations

These methods work extracting particular examples from the model to explain the behaviour of the ML model. They are similar to agnostic models but differs from these ones because they do not modify features or model. The techniques available in this approach are the *prototypes and criticism*, that overgeneralize the model extrapolating groups of similar instances called prototypes, that are well represented by the model, and at the same time extract also a group of instances badly represented, called criticism in order to do not overestimate the explanations given; the other one is the *counterfactuals explanations* which interestingly turn the point of view of all the techniques seen before: instead of explaining how the model generate the outputs, it studies which is the minimum condition that would lead to an output variation, so it focus on a single prediction against a group of different ones, like the work in [48].

At the end of this path, we can draw some conclusions: first we can say that the ways in which we can implement explainability are numerous and very varied. We have

seen three main methods to identify the different models, thus being able to adapt the choices on the particular situation in front of us. It must be said also that the challenges to improve the explainability of AI is still long and winding, and in practice it has to clash with several issues: first of all, the usual trade-off between interpretation and accuracy, which severely limits research in this area, as the tendency is to prefer methods that work well although not quite understandable.

another big problem is related to the profit that individual private companies have in the field of AI: algorithms created by private individuals are obviously monetized, so companies have no profit in making their models more transparent, because they would lose the uniqueness of their ideas. Moreover, the price of a program is obviously linked to its ability to perform a certain task as accurately as possible: this means that to make their model more interpretable, they should pay this with a loss in efficiency, which would obviously affect the price of the final product.

However, in the next chapter we will talk about the diffusion of the XAI in the field of interest of this work, so the medical one. Several algorithms are under research, and some are already implemented in some clinical areas, as the diffusion of AI still continuous.



## 4. *XAI in medicine*

Medicine is one of those fields that need special requirements, even for issues that may seem irrelevant if taken in the normal everyday life. The fragility of people entering medical processes, the protection of personnel who face dangerous situations for themselves and for patients, the structures that possess different degrees of danger according to the departments, with machinery that can also be very risky: everything that revolves around the health department requires stringent requirements to ensure reliability to the whole system.

An example of how stringent medical regulations are, is the definition of a medical device which is given in the European Directive 93/42/CEE which states that a medical device is “any instrument, apparatus, implant, substance or other product, used alone or in combination, including computer software used for proper operation and intended by the manufacturer for human use for: diagnosis, prevention, control, treatment or alleviation of a disease;...” [49]. The directive also defines software as medical devices, so that AI technologies also fall under this name, and must follow the same laws as these devices.

In addition to the purely developmental aspects, in medicine ethical and legal aspects are also fundamental, which seek to protect both the personnel who must use these devices, and the patients who are treated: strict laws and directives regulate any activity within the health care system, which protect people who are involved in it from different points of view such as the danger of environments and equipment, respect for the privacy of patients, the ethical nature of the decisions taken for surgical operations, diagnostic techniques and visits in general.

For all these reasons, the use of technologies that are not fully capable of being understood, and in a sense controlled, collides with the rigidity of all the regulations that are present in the medical sector. Think, for example, of the long and tortuous process necessary for the approval of a therapeutic plan or a medicinal product: years of studies to fully understand what the effects on people may be, to establish the real effectiveness of the product and so on. All these considerations have led the scientific world to promote research and development of AI technologies that are more

comprehensible and their applications in the healthcare system is becoming increasingly in demand. In fact, the level of knowledge offered by the XAI, allows to improve both the quality standards of the medical sector, as the staff is more aware of the equipment that uses, and at the same time the patient can better understand the results of his visits.

Another reason for the use of these technologies arises from the quality of data that we have available to train automated algorithms: only in recent decades the collection of information has evolved into a fully digital, while in previous years all the material was collected in paper documents. This obviously requires an immense amount of work from the developers who have to convert this documentation into a digital format, but even more problems arise from the quality of this information: in fact, the data currently available are strongly characterized by the presence of biases due to the behaviour of society in the past, which tended to exclude from social practices minorities, different ethnic groups and so on.

Being the information of the databases coming also from a past epoch, these are influenced by the presence of abundant data on a determined part of the population, typically of white ethnicity, while for all the other ethnicities the data are insufficient: if, for instance, an AI system is trained using a database that contain information from the USA government collected before the civil rights movement, “an algorithm ‘learns’ to prioritise patients, and it predicts to have better outcomes for a particular disease. This turns out to have a discriminatory effect on people within the Black and minority ethnic communities” [50]. Algorithms trained on such biased datasets could make considerably poorer predictions for, for example, younger black women [51].

If XAI became a routine in medical processes, these types of errors could be found and corrected, as the explanations provided by the system could direct developers to modify the training datasets improving not only the overall accuracy, but above all the ability to expand this accuracy to a larger pool of people.

Given all these critical aspects, there are several works in the literature that seek to implement XAI in medicine.

According to [52], the main clinical field that is in need for these technologies is the Clinical Decision Support Systems (CDSSs), so “systems support medical

practitioners in their clinic decision-making and in the absence of explainability may lead to issues of under or overreliance”. Always [52] identifies two main strands that exploit this technology: Image-based CDSSs and Linguistic Reasoners and Ontology-based CDSSs. The first deals mainly with algorithms for image analysis, with explanations provided in the form of user interfaces or emphasizing the most important portions of the original images; the other group focuses on more tabular and written information, by extracting the key concepts that led to the results.

The design of a CDSS system, must follow the indications of physicians, as they are the final users of these technologies. In the literature [53] [54], you can find studies in which you ask doctors who are used to use these tools to define what are, according to their experience, the basic characteristics that automated systems should have.

One of these attributes is the generation of explanations that show which is the correlations between the different features, and how they’re chained to produce the final decision: in this way the doctor not only compares the decision made by the automated system with his but can also compare the path made by the machine with what led him to his decision. This could be helpful not only in the single decision, but also for the experience of the doctor that can understand a different way to make a diagnosis.

Another characteristic is that physicians tend to be more confident with systems that justify their decision, even if they are less accurate than those that do not give explanations. It is obvious that a minimum accuracy threshold is necessary, to have a reliable machine, but for small changes in the efficiency of the systems, physicians prefer an explainable system in respect to a non-explainable one. The reason is that the machine is not giving a diagnosis, but it is assisting the doctor in the decision-making process, it is as if the automated system tries to promote or refute the decision that the doctor has already made through his experience: therefore, a physician tends to rely more on a result to which justifications follow, than on a simple final answer with no explanation behind it. So, the expert uses this machine in a more confident way, because even if the algorithm generates a not really reliable result, the physician can in any case analyse the reason behind this outcome and exploits this information to strengthen his decisions.

## 4.1 XAI application

In the following pages, some examples of XAI studies in medicine will be presented. As will be shown later, these technologies are also used in the most recent works: in particular two of the studies that will be presented, are focused on covid 19 and monoclonal antibodies, topics that in these moments are research trends.

### 4.1.1 COVID-19

The Covid-19 pandemic has radically changed the life of mankind, which has had to adapt to a situation not seen on earth since the spread of Spanish flu in 1918. Despite the current high level of technology, the virus has brought most countries to their knees, affecting in particular the economy and the health system: the latter in particular has found itself having to accommodate a very high number of patients, particularly in the first phase of the pandemic, both in its low-risk sections and in those of intensive care.

The physiological response of governments was that of restrictions and lockdowns: in the early period of spread, the lack of testing did not allow a screening of the population able to cope with the contagiousness of the virus, Therefore the only solution was to limit to the maximum the contacts between people, seen also the great presence of people asymptomatic, that did not give signs of contagion.

In this dramatic scenario, the search for alternative methods to quickly and efficiently diagnose the pathology was immediately activated, in order to find out in advance asymptomatic patients, or treat appropriately persons who were at an early and controllable stage of the disease before their condition worsened. Being a disease that primarily affects the respiratory tract, it was noted that radiological techniques, such as radiography or computed tomography, could be exploited as diagnostic imaging of Covid: in particular, classical radiography offers an almost immediate solution, with a relatively low cost and that under certain conditions can give an initial indication of the situation of patients.

In this branch of medicine, automated systems for image analysis can be efficiently integrated: there are already several studies on the subject and some models of artificial intelligence are already exploited abundantly. A particularly interesting study is that presented by Tsiknakis et al. [55]: these propose an algorithm for the analysis of chest X-rays that exploits a neural network architecture, focusing also on the interpretation of the system through the generation of heat maps on the images used as input.

The AI allows in fact to speed up the diagnosis of the infection from Covid and it can be thought to integrate this technology in the clinical practice, being able to replace the classic tampons in the situations of deficiency of the latter, or when a more immediate intervention is required, and therefore an early diagnosis. In addition, the high capacity of the neural network to extract even small details from the images, allows to detect initials and small signs of an early stage of the disease. If we add to all this the efforts made by the authors to make this system as explainable as possible, we are faced with a solution that certainly deserves the attention of the experts.

The work aims to recognize the presence of a SARS-Cov-2 infection by chest X-rays, which therefore include the lungs: the pathology in fact tends to create dark zones on these organs that in physiological situations should be clear to x-rays. The authors not only want to have results of the type “Covid vs healthy”, but also to distinguish the pathological cases between them: in fact, the possible outcomes can reach up to 4 different cases, that is “healthy” or “pneumonia due to covid” or “of bacterial” or “viral origin”.

The database used to train the system is composed by the union of existing datasets, from which 572 samples of posteroanterior view have been extracted, anonymized, and randomly mixed. These datasets [56] [57] [58] [59] include radiographic or tomographic images of healthy patients or patients suffering from pathologies indicated as possible results of analysis.

The number of samples contrasts with the information that has been provided so far in this thesis: in fact, about six hundred samples in input are a very small number, compared to the tens of thousands normally required by a system such as neural networks to be trained optimally. The reason that allows this small number of

samples is given by how the system is designed: in fact, the authors apply the so-called "transfer learning", so the generation of a complex and robust system trained on a huge general database, and then refine the algorithm capabilities with a second training phase, in which specific datasets are used for the desired application. In the literature it has already been shown that this technique succeeds within this work.

The architecture of the system reflects this strategy, as the algorithm is divided into two parts. The first is a convolutional neural network that is stripped of the final classification network, trained with a dataset of about 14 million images divided into different classes. The second is a classic neural network classifier consisting of a final level of 2 neurons (binary classification), 3 (ternary classification) or 4 (quaternary classification) and trained with the dataset of healthy and pathological patients. The results obtained are evaluated through a set of parameters: accuracy (ACC), sensitivity (SEN), precision (PRE), and area under the curve (AUC) for binary classification, to which are added the "AUC against all" or "against one" for multiclass classification. The results are extremely satisfactory, especially for the binary classification of "healthy vs covid" patients, reaching an accuracy of around 100%, and, in general, values that balance or exceed the current state of the art in all observation parameters. The results of the ternary classification (healthy vs COVID vs bacteria pneumonia) are less satisfactory because they do not reach the values of other works. Finally, for the quaternary one (healthy vs COVID vs bacteria pneumonia vs virus pneumonia) there are no reference parameters in literature, but in general it can be said that, although with a slightly lower accuracy than the others ( $76\pm 8\%$ ), the values of SEN, PRE, and so on are satisfactory although a bit unstable ( $93\pm 9$  for SEN,  $91.8\pm 7.6$  for PRE, and so on)

To make the algorithm explainable was implemented a model of interpretability called GradCAM [60]: this algorithm is able to follow the gradient variations between the first layer, the input one, and the final classification one. We then analyse the variations of the gradient of the score assigned to a particular class. Analysing this gradient, the system is able to generate a heat map of pixels that have contributed in priority to the generation of the specific output, then the class assigned to the image. Heatmaps are not related to the outcome of the classification, but only to the pathway followed by the CNN: for this reason, it is necessary to be careful not to interpret this information incorrectly, as the heat map is not indicating which of the

possible results is the correct one, but only where the network has identified the most important features to get to the result, whatever it is. For these reasons, it is possible to define this kind of interpretability as a local type, being each heat maps the description of a single analysis.

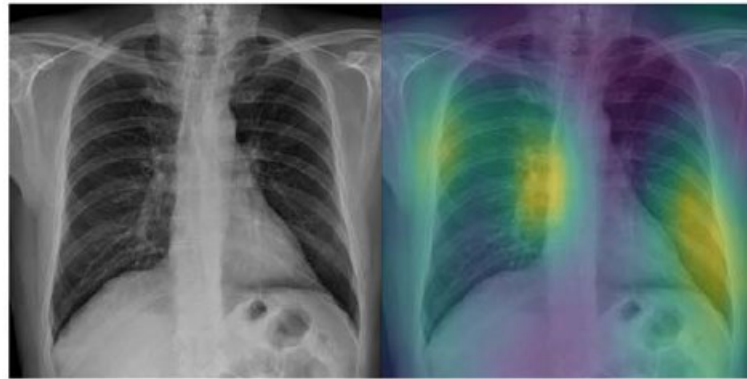


Figure 31: example of heat map

In Figure 31 is shown one example of one input image and the heat map generated by the system: the image refers to a binary classification problem, was taken by a patient positive to Covid and the model correctly identified its condition. The heat map shows what parts of the radiography the algorithm considers important, but nothing can be said on the result just looking at it.

In order to increase the reliability of the explanations generated by the system, specialized radiologists were questioned: they were asked to give an assessment to heat maps, to understand whether what the system considers important is really useful in order to correctly diagnose the condition of the patient under examination, so the lungs. They are asked to evaluate each hemithorax (about half image) with a score from 0 to 4, following these indications, taken directly from the paper:

0. The attention map is mostly homogeneous across the entire image
1. The attention map is focusing on totally irrelevant areas outside the lung
2. The attention map is focusing on the lung areas but also on other extrapulmonary structures
3. The attention map is focusing mostly on the lung areas
4. The attention map is focusing exclusively on the lung areas

The concentration of the network in areas outside the lungs (the organs attacked by the disease) is considered by the authors as a sign of lack of robustness and confusion. It is for this reason that radiologists have been interviewed: if the system is able to generate explanations on how it arrived at the result, an expert eye is able to judge this work, making it possible to understand where the system is wrong, so that we can then work on improving the algorithm. Indeed, one of the reasons identified for the misclassifications in ternary and quaternary problems is the need of larger dataset to train the classification net.

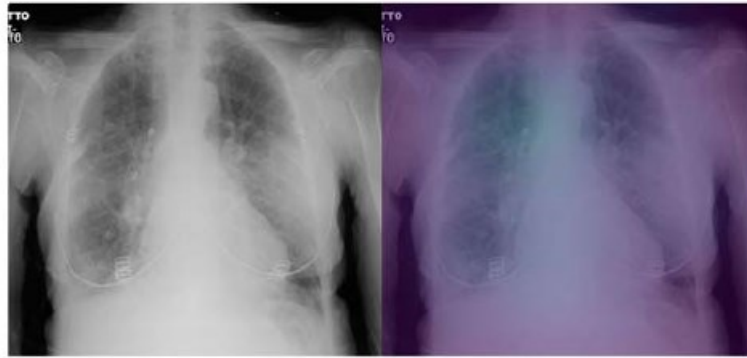


Figure 32: example of low and high level of explanations

An example of how the results and the heat maps are uncorrelated is given by Figure 31 and Figure 32: both the patients were positive to COVID, and both were classified into the correct category, the first in a binary problem, the second into a ternary one. but, while the first image was ranked by experts with a score of 3 for both hemithorax, the second took a score of 0 for the left lung (to the right in the image, as mirrored) and a 4 for the right lung (then to the left). This result is a demonstration of how the process leading to the generation of heat maps is completely unrelated to that of classification.



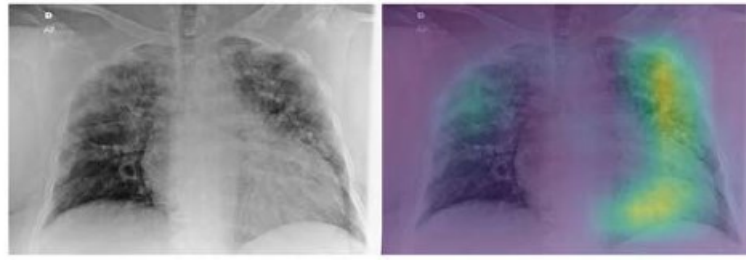


Figure 33: false negative patients

For instance, in Figure 33 the situation of a false negative patient in a ternary problem is shown, so the system classified the image as pneumonia while the correct result should have been Covid. Despite the misclassification, the experts gave a score of 3 and 4 respectively for the left and right lungs. This result is interesting, as it shows that, despite the classification errors, the system is extracting the fundamental characteristics from the lungs, even though the presence of a large number of details presents around the organs of interest. In this way it is possible to say that the problem is in the classification network, that is not trained optimally even if the previous part of the model is looking to the correct details.

This study certainly needs time to fully understand the mechanisms that can be improved to increase the efficiency of this model, especially regarding ternary and quaternary classification problems. At the same time, however, the results of the binary problem allow to apply this type of system also to other pathological cases, in particular thanks to its interpretability, which allows to understand how the network behaves in the particular application case, so the physicians can evaluate if the result of the automated system is valid or not.

#### 4.1.2 Monoclonal antibodies

Monoclonal antibodies are proteins created in the laboratory that can replicate the behaviour of our immune system, in particular antibodies [61]. The first to be approved for clinical use by the Food and Drugs Administration (FDA) in the United States was an antibody able to avoid rejection of a transplanted organ in case the

latter was resistant to corticosteroids [62]. The particular feature of this technique is the possibility of selectively developing antibodies that can focus on a specific target of the antigen to be addressed, producing on a large scale different monoclonal antibodies for each of the possible immune response mechanisms.

There are currently hundreds of studies that seek to improve existing therapies, or try to apply this technology in new areas, particularly in the treatment of immunological or oncological diseases. Even during the current COVID pandemic, the use of monoclonal antibodies was proposed, and in August also the AIFA (Agenzia Italiana del Farmaco) approved the use of this therapy [63] in several practical cases.

The process leading to the generation of an artificial protein, including monoclonal antibodies, is very complicated: the necessary resources are enormous, in terms of time, investment, articulated techniques and the need for special equipment. In particular, this technology usually starts with a huge population of potential proteins, which are gradually discarded if not working: the variables that can influence the efficiency of these products are many. For this reason, in recent years, research is focusing on trying to implement AI algorithms, to accelerate these inherently long processes.

In this context fits a very interesting work that has as authors Gentiluomo et al. [64]: they try to implement an explainable ANN to predict biophysical properties of therapeutic monoclonal antibodies, like melting temperature, aggregation onset temperature or interaction parameter, as a function of pH and salt concentration from the amino acid composition. The explainability of the systema comes from the number of input features, kept voluntarily low, and from the application of a model called “knowledge transfer”, whose aim is to understand how the ANN obtained its outcomes.

Their work starts from the desire to make more efficient and rapid the selection process of those monoclonal antibodies that can truly be studied for a future practical application. Imagine, for example, that usually the process of developing a protein drug starts with a population of candidates equal to a few thousand, and then get from these only 8% of drugs that will get a license. Under this condition, their idea is to apply an ANN model to extract the correlation of some important chemical and physical parameter of proteins, like the melting temperature, from the amino acid

composition of the antibodies, but as a function of basic characteristics like the pH and the ionic strength: the collection and the interactions of these characteristics, even if not effectively correlated to the stability of the products, can be useful to eliminate or validate a candidate of the study, knowing some basics rules from chemistry.

The primary goal of the authors was to have a network able to generate the desired outputs while maintaining a high level of explainability, in the sense of being able to understand how the network produced its results. For this reason, the choice of inputs has been limited to the composition of proteins in terms of amino acids, which allows to keep the model simple but at the same time accurate, even if the addition of further information could increase the modelling of reality. In addition to this, a second approach called "knowledge transfer" has also been used to understand the decision-making process of the algorithm.

The dataset is composed by 144 records (24 conditions per protein) for each parameter to be studied. The network consists in a feed-forward back-propagation network with one hidden layer of 5 neurons, therefore a relatively simple structure, compared for example to the topology presented for the application on radiographic images illustrated before.

The explainability of the system is obtained through the knowledge transfer, that is the creation of a second simpler algorithm, that allows to approximate the main ANN, and is resumed in Figure 34. The method begins with a study phase of the main algorithm after the training phase: it was noticed that none of the weights of the hidden level was zero, so all the nodes contribute to the result, but only 5% of the final values were at least twice as large as the average of the entire population. From this 5%, input parameters that have activation function coefficients at least double the mean of all values are extracted.

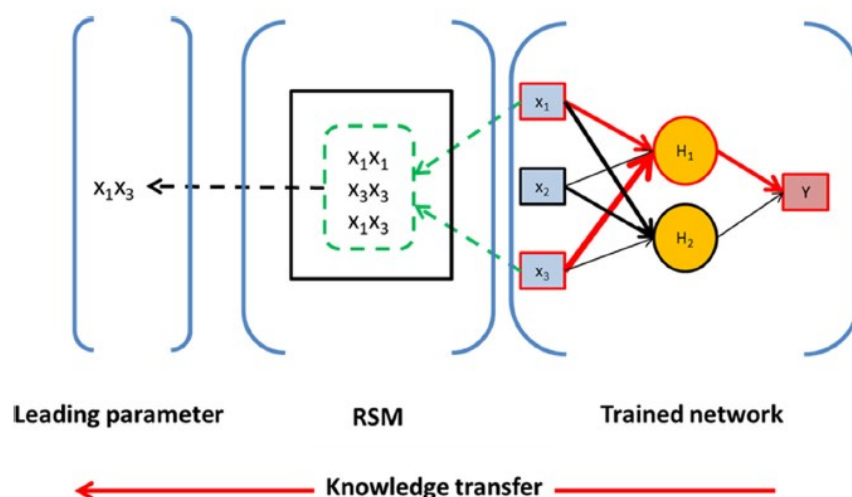


Figure 34: knowledge transfer process

Now that some parameters are available, a method called Response Surface Methodology (RMS), a statistic algorithm which explores the relationships between several explanatory variables and one or more response variables [65], is applied on the linear least squared regression of these extracted parameters. Finally, after reducing the model to terms statistically relevant, a curved response was obtained.

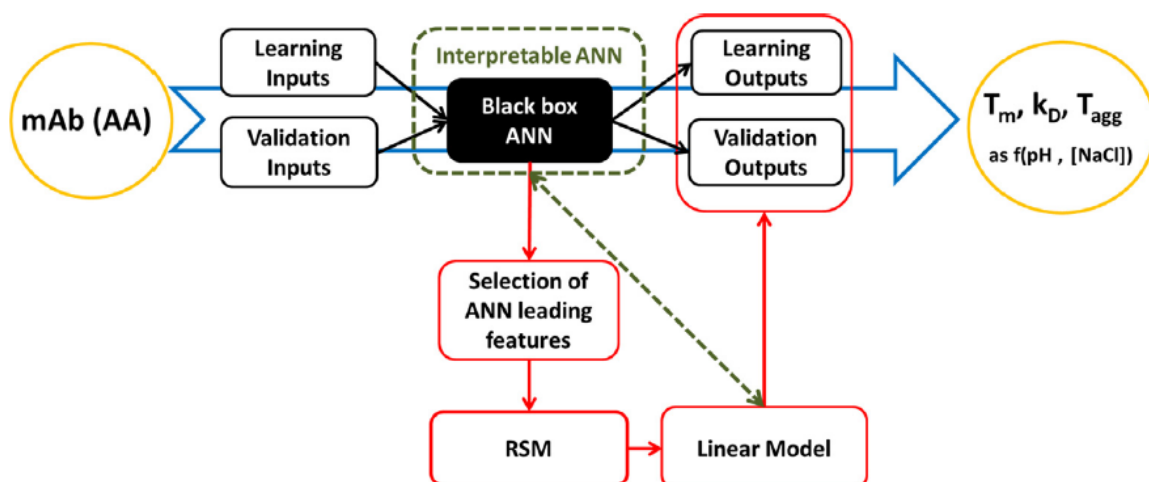


Figure 35: scheme of the interpretability approach

A scheme of this approach is given in Figure 35: in the image you can see that the predictive and explanatory process are not completely disjointed: in fact, building the explainability starting from the results, the explanations offered by the linear model are real justifications of the outputs. This differs from the previously proposed algorithm for COVID, in which interpretation and results were completely separate.

The main feature of this way to explain the network behaviour, in addition to the intrinsic simplicity of the model, is that it is possible to access the hidden layer of the ANN and to understand how the output and the input are related, allowing to easily retrain the algorithm and evaluate the explanation. On the other hand, this approach makes the explanation heavily linked with the training set, losing generalization.

From a practical point of view, the authors have been able to understand how the network estimates the parameters of interest once an acceptable accuracy value has been reached. For example, in the case of melting temperature, the main characteristic influencing the values is pH, saline concentration and the number of certain particular residues, including cysteine, which stabilizes the protein structure and consequently will influence the stability of monoclonal antibodies.

Correlations such as these have been extracted from all parameters of interest. In this way it is possible to have a general perception of the model, increasing the confidence of using the model that is no longer a black box. So, the strength of this application, is that the algorithm is able not only to produce numerical predictions of parameters, but also to detect the interactions between the different characteristics of these compounds, increasing general knowledge about the phenomenon and improving reliability and confidence of the model.

### 4.1.3 Heart failure

In emergency situations with patients with acute shortness of breath, one of the indicators of the occurrence of heart failure is the presence in high quantities of B-type natriuretic peptide (BNP). If the patient has normal values of this peptide, the presence of heart problems can be ruled out, while high values of this substance, in the presence of other determinants, may lead to diagnose a heart failure in time to

treat the disease. The main limitation of this technique is that the laboratory test of this compound is not often offered by all hospitals and readings of the test are not immediate.

Another method of diagnosing heart failure is that of chest radiography. We then look for a series of anatomical modifications visible with this imaging technique: examples are the pleural effusion, so the accumulation of fluid inside the pleura, cardiomegaly, so the enlargement of the cardiac silhouette, or Kerley B lines, horizontal lines in the periphery of the lower posterior lung fields, and others [66]. The main issue of this technique is the difficulty of reading this type of radiographs, not simple even for experienced radiologists: in this case the quality and timeliness of diagnosis is highly dependent on the degree of experience of the expert.

Seah et al. [67] propose a Generative Adversarial Network (GAN) based approach and the use of autoencoders to classify chest radiographs in search of the main signs of heart failure. In addition to the classification, the system also provides explanations in the form of images: the algorithm generates from the initial image a new X-ray, which represents how it should have been the original image to belong to the other available category. For example, if the input is classified as "healthy", the system justifies its solution by generating a new image in which it adds to the initial one the signs that are typically associated with the "pathological" label.

The GAN is a kind of ANN in which two different nets are trained in a competitive way, as in a one vs one game. The two opponents are two networks defined "generator" and "discriminator": the first aims to learn how to generate new data, learning the structure from a training dataset, while the second aims to be able to discriminate which data comes from the dataset and which were generated by the other network.

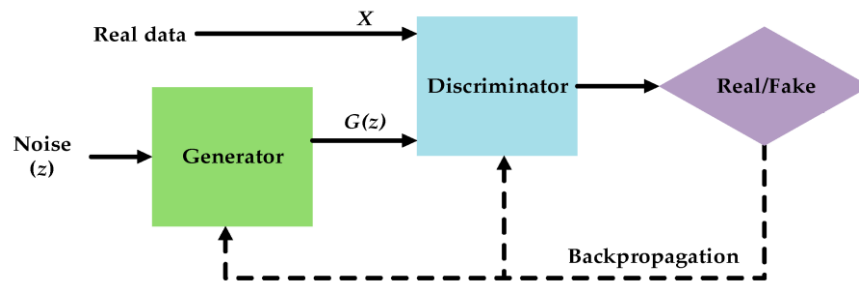


Figure 36: example of GAN

The autoencoders are a type of ANN used in unsupervised learning, whose aim is to learn a representation (encoding) of a dataset. This structure is composed by two parts, called encoder and decoder: the first maps the input into the “code”, so the mapping of the most important features of the dataset into a lower dimensionality set, while the decoder validates the mapping procedure trying to reconstruct the data from the code and comparing this information with the original set.

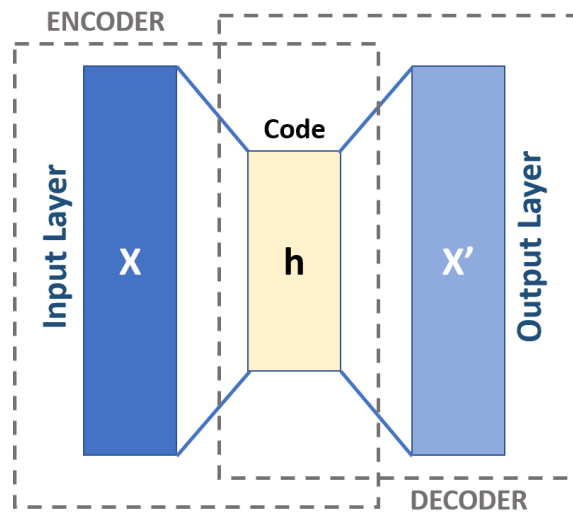


Figure 37: example of an autoencoder network

The authors, starting from the basic ideas of the two algorithms indicated, propose an explanatory model called *generative visual rationales (GVRs)*, which explain individual predictions, starting from a GAN approach.

The process is composed by three steps: first they build a GAN in which they train a generator through an unlabelled dataset, then they exploit this generator as the decoder of an autoencoder network, and finally they create a supervised model on the encoded representation of a small, labelled dataset. The explanation process is built implementing an optimization problem that aims to change the predicted class of the input, but at the same time penalizing the differences between the new image and the original one. This process generates images similar to the original ones, but to which are removed or added the details necessary to be classified in the category opposite to the one assigned to them.

The databased used for the training of the GAN is composed by around 90 thousand chest radiographies; then they train the autoencoder to perfectly reproduce the dataset images using, for each patient in the database, 15 images as training set and 1 for validation test. Finally, a regressor was trained using a labelled database of around seven thousand chest radiographies paired with a BNP blood test value, which we said was linked to heart failure.

This process allows to generate an image identical to the original, but with the details that would have led to a different classification: overlapping or subtracting the images you can get the GVR explanation for each image. To demonstrate the behaviour of the system, it was applied to a set of imaged not seen in the training phase. The model achieved an area under the curve (AUC) of 0.837 using a linear regressor.



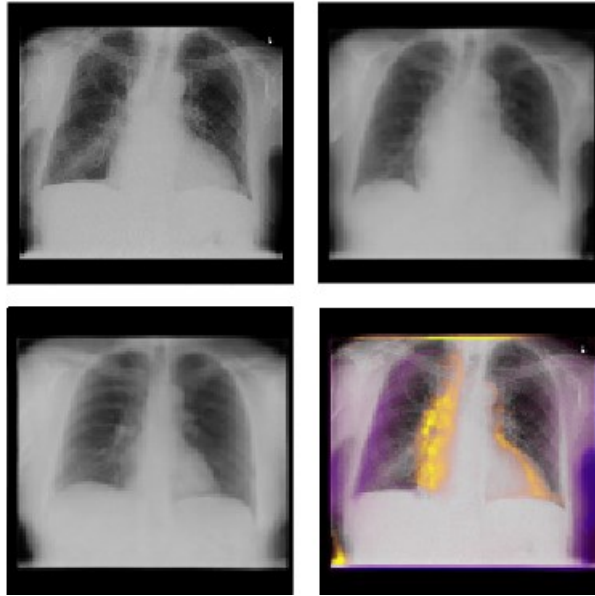


Figure 38: example of GVR

In Figure 38 is shown an example of the GVR on a chest radiography of a patient with heart failure: upper left is shown the original X-ray of the patient, while on the right the one generated by the system without any sign of heart failure. At the bottom left we have the original image again, to compare it with the right one in which are indicated the areas that the algorithm observes to classify the image. To achieve this result the algorithm reconstruct the image of the pathological patient but removing the signs of heart failure like the pleural expansion on the right lung and the increase in heart volume on the left lung. Then the model subtracts the non-pathological image from the original one generating a heat map, in which positive values are displayed as purple and negative as orange, which is superimposed to the input image demonstrating the GVR for its prediction.

In this work is shown that autoencoders can be exploit in this medical field, and in particular that they can be used to produce not only an acceptably accurate prediction, but also to generate explanations of their decision through the GVR model. Of course, some limitations are present: the authors indicate that there is not a rigorous definition of what interpretability is, and they proposed that one metrics could be how much explanation would need a second model to reach the same accuracy by learning from this interpretation; in this way would be possible to train different algorithms with smaller database, transferring the already known information

extracted by the first model. Finally, a technical problem is that the networks of this work struggle to produce images bigger than 128 by 128 pixels, which is not a perfect resolution to extract small details from images.

#### 4.1.4 Breast cancer diagnosis

Cancer is currently the disease that causes the most deaths in the world: in 2020 alone, there were about 19 million diagnoses and 10 million deaths. According to the Global Cancer Statistics 2020 [68], one in five people will develop a cancer in their lifetime. If we focus on women, breast cancer is the most common invasive cancer in women and the second leading cause of cancer death in women.

For these reasons in the last years the number of screening processes increased a lot, in particular in developed country, exploiting the speed and ease of implementation of radiological techniques such as mammography. However, the analysis of these images requires a meticulous effort on the part of experts, in search of spots or nodules, which must take a long time in the observation of X-rays. Moreover, the ability to diagnose depends on the quality of the images, the visual skills of the operator and his experience.

In this context, the use of automated techniques for image analysis can be easily inserted, which allow to speed up this process and to assist doctors in their diagnosis work. There are already several studies that are applied, but they usually come in the form of black-box models, which as widely said are not viewed favourably by the medical environment, as they lack transparency.

Interesting is the work done by Brito-Sarracino et al. [69]: Their aim is to create an automated system of ML that allows to obtain an accurate diagnosis of breast cancer, maintaining a high level of explainability, not only with regard to the classification phase, but also throughout the phase of feature selection. In fact, not only do they use an interpretable approach to discriminate against the presence or not of a tumour within an X-ray image, but with a visualization process they allow to have a high degree of interpretability also in the selection of the main characteristics of

discrimination. The paper concentrates its attention on the discrimination between benign (B) and malignant (M) breast cancers.

They utilize a dataset obtained by the University of Wisconsin Hospitals [70], which consists of 569 observations and 31 variables, including the target variable, diagnosis, and excluding the ID number. The entire set of features are calculated based on 10 principal features, that mainly describe the tumoral cell nuclei. The dataset is then split into a training set (80% of records) and test set (20%).

The feature selection phase is done exploiting two different techniques, called *visualization techniques* and *Recursive Feature Elimination (RFE) wrapper*. The first one selects the top scoring attributes, while the second approach takes as input not only the features, but also a feature relevance estimator model, a set of labelled examples and the number of features to be selected: from these inputs it iteratively discards one feature at time, choosing the one that reach the worst results from the estimator model. In this phase the authors implement explanation of their results through two explanatory models, called *Linear Projections (LP)* and *Radviz (RV)*. These approaches of data visualization are very useful, in particular because they allow to increase the level of knowledge of the correlations between input and output, and allow to perceive in a more direct way which are the classes in which the records are separated

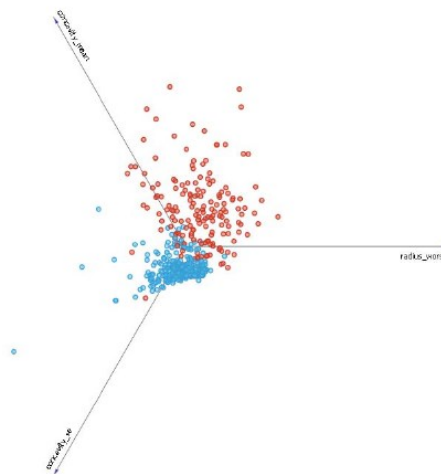


Figure 39: example of LP feature visualization

The LP method is a sort of representation of the distribution of records within an axis diagram, in which each of them represents one of the main features, like the one in Figure 39. In this figure the three axes represent the worst mean distance of point from the centre of the image, the mean concavity and concavity standard error. The colour of records represents their class, in red the benign class, in blue the malignant one. It is possible also to increase the number of the axes.

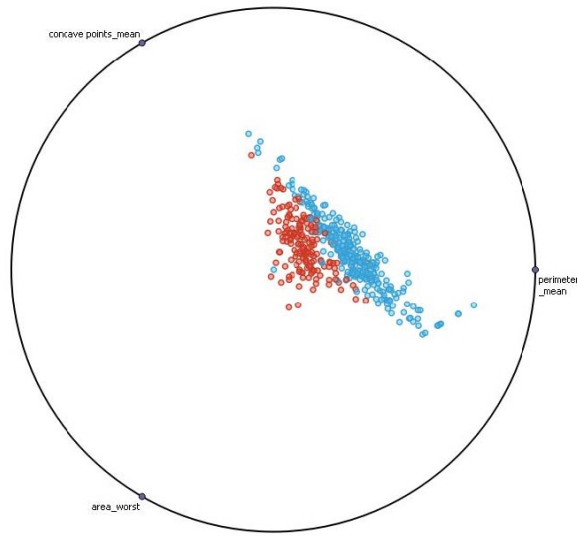


Figure 40: example of Radviz visualization

The second approach is the Radviz one: in this method the features are arranged on a circumference, while the points are arranged in the inner circle. The dependence on a certain variable is represented by the tendency of the points to approach the part of circumference on which lies the particular feature. The radius of the circle is unitary, so the points are in a position proportional to the real attribute values that span in a range between 0 and 1. Also in this case the number of variables can increase.

Finally, the classification model chosen was the *Classification and Regression Trees* (CART), supported by a technique called Grid Search [71], able to extract the most important hyper-parameters. The CART was used in different ways: one does not use feature selection or visualization (baseline), three select 3 features using LP(3),

RV(3) and RFE(3), another three select 5 features through LP(5), RV(5), RFE(5) and finally one uses cross validation for independently choose the number of features to be extracted through RFE(CV). Results are shown in Figure 41.

Classifier	Accuracy	Precision	Recall	F1-score
baseline	0.90	0.88	0.86	0.87
RFE(CV)	<b>0.96</b>	0.93	0.98	<b>0.95</b>
RFE(3)	<b>0.96</b>	<b>0.95</b>	0.93	0.94
RFE(5)	0.92	0.88	0.90	0.89
LP(3)	<b>0.96</b>	0.89	<b>1.00</b>	0.94
LP(5)	0.89	0.89	0.81	0.85
RV(3)	<b>0.96</b>	0.93	0.95	0.94
RV(5)	0.92	0.82	<b>1.00</b>	0.90

Figure 41: results of the classifiers

It must be notice that, even if the CART used as baseline is the worst one in terms of results, it is the most interpretable one, as its structure, depicted in is simpler than all the other ones.

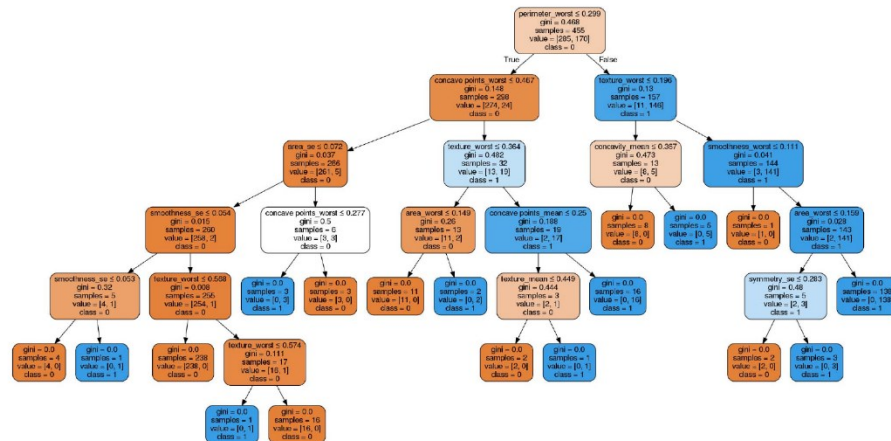


Figure 42: baseline CART structure

The Wilcoxon test was used to determine whether there were statistical differences between using the CART with RFE or with visualization techniques, using a p-value  $< 0.5$  as the separation threshold. The results obtained shows that there is no difference between the two approaches: this is an important outcome, as it is possible to apply visualization techniques without losing accuracy, allowing to understand the reasons that lead the system to discern between different classes, bringing to light correlations that are not easy to find through human analysis or black-box models. Indeed, even if other approaches can reach better accuracy values, like for instance ANNs, their opacity make really difficult to understand which are the important parameters that drive their decisions.

## **4.2 Limitations and challenges**

So far, XAI has been discussed from a purely practical point of view, presenting the state of the art of this technology and in particular its application within a particular sector, the medical one. At present, this branch of research is still at a very early stage: there are many problems that need to be addressed, but at the same time the potential of this approach is very high.

Within this chapter we have tried to present as much as possible the improvements offered by explainable technology, compared to the state of the art of classic AI. This, however, could be misleading, as you might be led to think that the XAI are an evolution of ML techniques, which solve some limitations of these algorithms without any kind of problem. In practice, there are major problems that need to be tackled, even more on the assumption that this sector is at an early stage of full development. In response to these strong limitations, however, several potential replies to these limitations are already under research, which could allow an even greater diffusion of these powerful tools, capable of leading to technological evolution and greater digitalisation of different sectors.

However, if we want to deal with the limitations and possible future improvements for XAI, it is necessary to consider that this method is strongly influenced by the application environment in which you work: as we have already said, each sector has

a greater or lesser need of the explainability of these systems, and this depends on the degree of responsibility that weighs on the actions of these machines.

The evolutionary process of XAI systems is characterized by the close relationship between limitations and potential, as they are the first to give strength to the development of the latter. It becomes therefore complicated to analyse in a completely independent way which are the problems and which the directions of future development, because it is natural to face the two issues in parallel.

One of the main drawbacks derives mainly from the lack of formalism that characterizes the literature on this subject. In fact, being a relatively young field of research, the authors tend to behave as single entities that work by watertight compartments, that is moving within their comfort zones, trying to maximize their results in the absolute sense, without comparing it to other similar works.

This phenomenon is due to the fact that there are no formal concepts for the different characteristics of these systems: for example, a formal concept of explainability is not available, whose description remains anchored to the extreme variability with which the problem can be addressed, also from the human point of view. In fact, if you wanted to give a unique definition of what is explainability in daily life, the possible answers would probably all be different, because of the infinite nuances that can assume: explain something to someone remains an extremely personal action, influenced by the social context in which you find yourself, by the interlocutor with whom you approach, by the level of finesse required to the explanation to be provided.

Precisely these differences are also the cause of another important limitation, which is that of how to evaluate the interpretability. Not existing a formal definition, it is not possible to develop univocal metrics that allow to define how much a system is interpretable. When it comes to AI in decision-making, we can navigate through the multiple units of measurement available to define the quality of an algorithm: accuracy in results, sensitivity to a particular aspect of the analysis, or mathematical concepts such the area under the curve. The same cannot be said for the concept of explainability, for which there is no metric that allows comparisons between XAI systems, even if they face the problem with similar, if not equal, approaches.

In this situation of uncertainty, it is physiological that the different authors try to give an evaluation to their work with various approaches to the problem: in [72] are mainly identified three types of approaches for evaluation of methods, namely *application-grounded*, *human-grounded* and *functionally-grounded*. In the first case, the measure of the interpretability must be done at the level of the practical application of the AI system, entrusting the evaluation to an expert in the field, then the end user of the algorithm. The *human-grounded* approach, on the other hand, aims to measure the explainability of a system through the eye of people who are not experts in the sector, so we try to make more general judgments on the quality of the explanations provided, rather than study how well they are suited to understanding the specific field of application. The last method does not involve humans but tries to compare and regularize the different methods after the results have already been validated.

In a context such as the medical one, in which the delicacy of the applications reaches extremes difficult to compare with other sectors, the need to compare the different methods becomes fundamental. In any health care system, no aspect is left to chance, and even issues of minimal relevance, are constantly kept under control, in order to make any environment safe for the staff and the patient and to offer the highest possible efficiency, which is reflected in the quality and speed of diagnosis and accuracy of treatments. In this meticulous process of control, automated systems of medical assistance, in particular those of decision-making, which are medical devices to all intents and purposes, cannot be exempted, as previously indicated.

For these reasons, it is necessary in the short term to formalize all aspects that revolve around the XAI sector: in doing so it will be possible to define the evaluation fees of these instruments, allowing comparisons to be made, to judge the different models proposed, and more generally to regulate in a univocal way the different particularities of this technology.

Another very important problem that afflicts the XAI is the reproducibility of their behaviour: as we have seen also in the examples proposed in the previous paragraph, the generation of these systems is strongly linked to the specific application in which an algorithm is inserted. The explanations that are generated in a post-hoc way, are usually deeply dependent on the main system, which is conceived and trained on the



basis of particular sets of information and with a specific final objective: in this contest it becomes extremely complicated to create explanatory models completely adaptable to algorithms other than native ones, as the transparent XAI models try to do. The level of generalisation is therefore not high, and this leads to problems not only at a practical level, but above all at the level of the economic resources and time needed to differentiate different models that can be explained to all possible applications.

Finally, we cannot exclude the fact that the XAIs, like any other instrument, are subject to economic factors deriving from private companies that develop these types of systems. It is obvious that companies dealing with software development or creating biomedical devices work to generate a profit, whether this is put in the first or second place with respect to the goal of global technological development: This creates a series of doubts that fit into wider contexts of legal, ethical, and economic issues.

From a legal point of view, it is not yet clear how the advent of automated technologies can be regulated: being technologies currently used mainly in decision-making assistance, and not in generate automated choices, it is possible to correct any errors, or in any case be excluded by the operator who has to make a diagnosis. This could change, in the event that these tools enter into clinical practice to make independent choices, as there would be a need for a very strict regulation, which imposes legal and ethical responsibilities in their use, especially in those situations where people's lives are based on the results of such algorithms.

Even more delicate in the case of XAI is the economic aspect, because private companies could not appreciate the spread of these technologies: the profit of these companies is linked to the quality of the systems introduced on the market, which therefore tend to shape reality as reliably and efficiently as possible. Typically, as has been said above, the best systems are the most complicated, which at the same time are the most opaque: making these algorithms more transparent would force companies to a choice in the context of trade-off between the accuracy of the systems and their interpretability. The possible way could be to make algorithms less complex gaining in explainability but paying for this choice with a less accurate analysis of the data. Otherwise, it is possible to maintain a high level of complexity but

increasing efforts to make the model more transparent: the latter choice is the most complicated for a company, which obviously has a gain from the originality of its algorithms, especially in those cases where there are patents that protect its uniqueness.

In the coming years, the scientific community will necessarily have to make choices, and make decisions about all these problems illustrated so far, which are just the tip of the iceberg of a more articulated system. Surely it will be necessary to integrate in the development of these systems different disciplines, both technical (computer science, engineering) but also more "human" (psychology, pedagogy). This integration is fundamental in the development of a formal theory, which would allow us to take a common direction in the technological evolution of these instruments. In addition, it will be necessary to involve in a massive way the experts in the sector, the specialized technicians, and more generally all end-users who will then have to evaluate, buy, and use the equipment in development.

In addition, there is the need to develop more specific legislation as soon as possible, without creating ambiguities or doubts about the development or use of these instruments: European directives such as the GDPR must only be the beginning of a phase of becoming aware of what direction the world is taking, which in the future will be increasingly pervaded and influenced by the technological advancement of automated systems. Laws, directives, insurance policies, everything that currently exists for issues such as work, economics, medicine, justice, must also be transported in the field of artificial intelligence, which is no more a "futuristic thing", but it is an actual and huge development sector.

## 5. *Conclusions*

The path addressed in this work tells us that although the AI sector is already at a very high level, the potential improvements are still a lot: in this evolutionary process the XAI are an excellent candidate, in particular for those sectors that have very stringent requirements and require a high degree of controllability of instruments. In these areas it is essential to have an alternative to black-box models, which allow to increase the level of transparency without losing the accuracy in data analysis that characterizes automated technologies.

As we have seen, the possibility of obtaining from AI systems explanations and justifications of their choices allows to increase the qualitative level not only of the instrument itself, but also of the entire decision-making system in general: in fact, using the XAI allows to improve the controllability of the instrument, to raise the level of confidence of the system in its field of application, to discover new knowledge understood as deeper correlations between different aspects of the phenomenon under analysis or new ways of exploiting the technology under consideration. Obviously, as this technology is in its infancy, a collective and common effort of the scientific community is still needed in order to improve a tool that is intended to revolutionize the use of automated systems, expanding the fields of application even to those most reluctant to accept them in common practice.

Major efforts should be made to involve industry experts as much as possible, as the XAI are perhaps one of the tools most in need of the opinions and views of end-users: this is because, as has been widely discussed above, the explanations that the system must be able to provide must be suitable for the precise recipient, thus having to adapt to the multitude of possible final interlocutors. Another objective should be the common and unambiguous definition of the requirements and metrics useful in the development of these technologies, which would make it possible to achieve an even better level of development.

However, we must also say a few words about the way in which the world looks at these instruments. The rate at which the world is currently evolving has never been so high: technological development over the past 50 years has progressed at a very

high rate, leading on the one hand to a general improvement in living and working conditions, But on the other hand it has generated a certain degree of mistrust towards an increasingly intrusive technology: it is necessary to place in this perspective the fear by some people of being "replaced" by machines. In the medical field in particular, it is necessary to work in the short term to the training of personnel who are able to manage and use in the best way all these automated tools, the purpose of which is not to replace, but to assist people in their work.

Reluctance is mainly a symptom of the inability of a generation of professionals to accept the instruments that are developed in recent years, considering them unable to balance the human skills of patient care. On the one hand, it cannot be denied that experience and intuition are fundamental characteristics in the medical field that machines cannot achieve: in fact, it is not uncommon for people's lives to be saved by the intuition of the doctor, despite all the data available did not support the expert's decision. These situations are not achievable by an automated system, which will seek the solution only and exclusively within its representation of reality. But it is precisely for this reason that professionals should consider these tools, because they can assist them in their decisions, allowing them to confirm their decisions, question them and maybe improve their level of confidence at work.

This improvement, however, requires a compromise: if it is true that the staff must be more flexible, it is necessary that even the developers of automated systems make an effort to make their technologies more suitable for the industry. Under this hypothesis the use of XAI can certainly speed up this process, allowing to find a middle way between the one required by the machine (accuracy, transparency and speed) and the one required by experts (updating, elasticity, renewal).

The study presented by this thesis shows how the current possibilities already allow a small first step forward, and that systems that can be implemented even now in practice are available. It is now up to the scientific community to pursue this project, with a view to a general improvement of the world, which will necessarily be increasingly invaded by AI technologies, but this process must be followed with conscience and criteria, in a constant balance between safety, control, accuracy and speed. Only in this way will we be able to derive all the possible benefits from the

technology (which are huge and numerous), avoiding that these tools are used in an reckless way.

# Bibliography

- [1] E. Commission, “Ethics guidelines for trustworthy AI,” 2019.
- [2] S. A. Bini, “Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care?,” *The journal of Arthroplasty*, 2018.
- [3] H. Salehi and R. Burgeño, “Emerging artificial intelligence methods in structural engineering,” *Engineering Structures*, no. 171, pp. 170-189, 2018.
- [4] M. Haenlein and A. Kaplan, “A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence,” *California Management Review, Special Issue on AI*, 2019.
- [5] T. Pavlidis, *Structural Pattern Recognition*, Springer-Verlag Berlin Heidelberg GmbH, 1977.
- [6] Quora, “What Exactly is Machine Intelligence,” Forbes, [Online]. Available: <https://www.forbes.com/sites/quora/2019/11/15/what-exactly-is-machine-intelligence/?sh=77462ca9187c>.
- [7] W.-L. Chao, “Introduction to Pattern Recognition”.
- [8] E. Commission, “A european strategy for data,” 2020.
- [9] C. Molnar, *Interpretable Machine Learning*, 2021.
- [10] C. Vercellis, *Business Intelligence*, Wiley, 2009.
- [11] P. Cunningham and M. Cord, *Machine Learning Techniques for Multimedia*, Springer, 2008.
- [12] Wikipedia, “Support-vector machine,” Wikimedia Foundation, [Online]. Available: [https://en.wikipedia.org/wiki/Support-vector\\_machine](https://en.wikipedia.org/wiki/Support-vector_machine).
- [13] Wikipedia, “Unsupervised Learning,” Wikimedia Foundation, [Online]. Available: [https://en.wikipedia.org/wiki/Unsupervised\\_learning](https://en.wikipedia.org/wiki/Unsupervised_learning).

- [14] N. Boldrini, “Deep Learning, cos'è l'apprendimento profondo, come funziona e quali sono i casi di applicazione,” 9 August 2019. [Online]. Available: <https://www.ai4business.it/intelligenza-artificiale/deep-learning/deep-learning-cose/>.
- [15] K. D. Foote, “A Brief History of Deep Learning,” 7 February 2017. [Online]. Available: <https://www.dataversity.net/brief-history-deep-learning/#>.
- [16] A. Wolfewicz, “Deep learning vs. machine learning - What's the difference?,” 3 May 2021. [Online]. Available: <https://levity.ai/blog/difference-machine-learning-deep-learning>.
- [17] Wikipedia, “Artificial Neural Network,” Wikimedia Foundation, [Online]. Available: [https://en.wikipedia.org/wiki/Artificial\\_neural\\_network](https://en.wikipedia.org/wiki/Artificial_neural_network).
- [18] V. Bansal, “The Evolution of Deep Learning,” 5 April 2020. [Online]. Available: <https://towardsdatascience.com/the-deep-history-of-deep-learning-3bebeb810fb2>.
- [19] Wikipedia, “Neural circuit,” Wikimedia Foundation, [Online]. Available: [https://en.wikipedia.org/wiki/Neural\\_circuit](https://en.wikipedia.org/wiki/Neural_circuit).
- [20] P. Cerveri, Dispense di Neuroengineering, 2019.
- [21] D. Ugochi, Y. Zhou, K. K. Deveerasetty and Q. Wu, “Unsupervised Learning Based On Artificial Neural,” in *2018 IEEE International Conference on Cyborg and Bionic Systems*, Shenzhen, 2018.
- [22] K. O'Shea and R. Nash, “An Introduction to Convolutional Neural Networks”.
- [23] S. Saha, “A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way,” towards data science, December 2018. [Online]. Available: <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53>.
- [24] R. Yamashita, M. Nishio, R. Kinh Gian Do and K. Togashi, “Convolutional neural networks: an overview,” *Insights into Imaging*, 2018.
- [25] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz and H. J. Aerts, “Artificial intelligence in radiology”.

- [26] S. R. Steinhubl and E. J. Topol, "Moving From Digitalization to Digitization in Cardiovascular Care: Why Is it Important, and What Could it Mean for Patients and Providers?," *Journal of the American College of Cardiology*, vol. 66, 2015.
- [27] D. L. Boeldt, N. E. Wineinger, J. Waalen, S. Gollamudi, A. Grossberg, S. R. Steinhubl, A. McCollister-Slipp, M. A. Rogers, C. Silvers and E. J. Topol, "How consumers and physicians view new medical technology: Comparative survey," *Journal of medical internet research*, vol. 17, no. 9, 2015.
- [28] K. Johnosn, J. T. Soto, B. Glicksberg, K. Shameer, R. Miotto, M. Ali, E. Ashley and J. Dudley, "Artificial Intelligence in Cardiology," *Journal of the American College of Cardiology*, vol. 71, no. 23, June 2018.
- [29] S. A. Halim, M. Neely, K. Pieper, S. Shah, W. Kraus, E. Hauser, R. Califf, C. Granger and L. Newby, "Simultaneous consideration of multiple candidate protein biomarkers for long-term risk for cardiovascular events," *Circulation: Cardiovascular Genetics*, February 2015.
- [30] S. Cui, K. Li, A. Lawrence, J. Liu, L. Cui, X. Song, S. Lv and E. Mahmud, "Plasma Phospholipids and Sphingolipids Identify Stent Restenosis After Percutaneous Coronary Intervention," *JACC: Cardiovascular Interventions*, vol. 10, no. 13, 2017.
- [31] K.-H. Yu, A. L. Beam and I. S. Kohane, "Artificial intelligence in healthcare," *Nature Biomedical Engineering*, no. 2, 2018.
- [32] B. Marr, "First FDA Approval For Clinical Cloud-Based Deep Learning In Healthcare," January 2017. [Online]. Available: <https://www.forbes.com/sites/bernardmarr/2017/01/20/first-fda-approval-for-clinical-cloud-based-deep-learning-in-healthcare/?sh=3d28e0ae161c>.
- [33] Arterys Inc., "ARTERYS," [Online]. Available: <https://arterys.com/>.
- [34] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, 2017.



- [35] G. Briganti and O. Le Moine, “Artificial Intelligence in Medicine: Today and Tomorrow,” *frontiers in medicine*, 2020.
- [36] M. Turek, “Explainable Artificial Intelligence (XAI),” DARPA, [Online]. Available: <https://www.darpa.mil/program/explainable-artificial-intelligence>.
- [37] A. Adadi and M. Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),” *IEEE Access*, vol. 6, 2018.
- [38] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf and G.-Z. Yang, “XAI—Explainable artificial intelligence,” *SCIENCE ROBOTICS*, 2019.
- [39] G. Montavon, W. Samek and K.-R. Muller, “Methods for Interpreting and Understanding Deep Neural Networks”.
- [40] P. Gohel, P. Singh and M. Mohanty, “Explainable AI: current status and future directions,” *IEEE Access*.
- [41] D. C. Elton, “Self-explaining AI as an alternative to interpretable AI”.
- [42] F. K. Dosilovic, M. Brcic and N. Hlupic, “Explainable Artificial Intelligence: A Survey,” in *MIPRO 2018*, Opatija, Croatia, 2018.
- [43] H. Lakkaraju, S. H. Bach and J. Leskovec, “Interpretable Decision Sets: A Joint Framework for Description and Prediction,” in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016.
- [44] R. Piltaver, M. Lustrek and M. Gams, “Multi-objective learning of accurate and comprehensible classifiers – a case study,” in *7th European Starting AI Researcher Symposium - STAIRS 2014*, Prague, Czech Republic, 2014.
- [45] S. Krening, B. Harrison, K. M. Feigh, C. L. Isbell, M. Riedl and A. Thomaz, “Learning From Explanations Using Sentiment and Advice in RL,” *IEEE Transactions on Cognitive and Developmental Systems*, vol. 9, no. 1, 2017.
- [46] A. Mahendran and A. Vedaldi, “Understanding Deep Image Representations by Inverting Them”.

- [47] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, “Distributed Representations of Words and Phrases”.
- [48] S. Wachter, B. Mittelstadt and C. Russell, “Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR,” *Harvard Journal of Law & Technology*, 2018.
- [49] Council of the European Communities, *COUNCIL DIRECTIVE 93/42/EEC*, 1993.
- [50] C. Garattini, J. Raffle, D. N. Aisyah and et al., “Big Data Analytics, Infectious Diseases and Associated Ethical Impacts,” *Philosophy & Technology*, vol. 32, 2019.
- [51] E. Vayena, A. Blasimme and G. I. Cohen, “Machine learning in medicine: Addressing ethical challenges,” *PLoS Med*, vol. 15, no. 11, 2018.
- [52] A. M. Antoniadis, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B. A. Becker and C. Mooney, “Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review,” *Applied Science*, vol. 11, no. 5088, 2021.
- [53] A. Bussone, S. Stumpf and D. O’Sullivan, “The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems,” in *2015 International Conference on Healthcare Informatics*, 2015.
- [54] S. Tonekaboni, S. Joshi, M. D. McCradden and A. Goldenberg, “What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use”.
- [55] N. Tsiknakis, E. Trivizakis, E. E. Vassalou, G. Z. Papadakis, D. A. Spandidos, T. Aristidis, J. Sanchez-Garcia, R. Lopez-Gonzalez, N. Papanikolaou, A. H. Karantanas and K. Maria, “Interpretable artificial intelligence framework for COVID-19 screening on chest X-rays,” *Experimental and therapeutic medicine*, vol. 20, 2020.
- [56] J. P. Cohen, P. Morrison and L. Dao, “COVID-19 Image Data Collection,” 2020.
- [57] D. Kermany, M. Goldbaum, W. Cai, C. Valentim, H. Liang, S. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. Prasadha, J. Pei, M. Ting, J. Zhu, C. Li, S.

- Hewett, J. Dong, I. Ziyar, A. Shi, ... and K. Zhang, “Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning”.
- [58] P. Mooney, “Chest X-ray images (Pneumonia),” [Online]. Available: <https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia>.
- [59] Quibim, “Covid-19,” [Online]. Available: <https://quibim.com/tag/covid-19/>.
- [60] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [61] Wikipedia, “Monoclonal antibody,” Wikimedia Foundation, [Online]. Available: [https://en.wikipedia.org/wiki/Monoclonal\\_antibody](https://en.wikipedia.org/wiki/Monoclonal_antibody).
- [62] Wikipedia, “Anticorpi monoclonali approvati per uso clinico,” Wikimedia Foundattion, [Online]. Available: [https://it.wikipedia.org/wiki/Anticorpi\\_monoclonali\\_approvati\\_per\\_uso\\_clinico](https://it.wikipedia.org/wiki/Anticorpi_monoclonali_approvati_per_uso_clinico).
- [63] Agenzia Italiana del Farmaco (AIFA), “Definizione delle modalità ottimali d’uso degli anticorpi monoclonali anti COVID-19,” 5 August 2021. [Online]. Available: <https://www.aifa.gov.it/-/definizione-delle-modalit%C3%A0-ottimali-d-uso-degli-anticorpi-monoclonali-anti-covid-19>.
- [64] L. Gentiluomo, D. Roessner, D. Augustijn, H. Svilenov, A. Kulakova, S. Mahapatra, G. Winter, W. Streicher, A. Rinnan, G. H. Peters, P. Harris and W. Frieß, “Application of interpretable artificial neural networks to early monoclonal antibodies development,” *European Journal of Pharmaceutics and Biopharmaceutics*, vol. 141, 2019.
- [65] Wikipedia, “Response surface methodology,” Wikimedia Foundation, [Online]. Available: [https://en.wikipedia.org/wiki/Response\\_surface\\_methodology](https://en.wikipedia.org/wiki/Response_surface_methodology).
- [66] B. Rasuli and J. Jones, “Heart failure (summary),” [Online]. Available: <https://radiopaedia.org/articles/heart-failure-summary>.
- [67] J. Seah, J. Tang, A. Kitchen and J. Seah, “Generative Visual Rationale”.

- [68] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal and F. Bray, “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” *CA: A Cancer Journal for Clinicians*, vol. 7, no. 3, pp. 209-249, 04 Feb 2021.
- [69] T. Brito-Sarracino, M. R. Santos, E. F. Antunes, I. B. A. Santos, J. C. Kasmanas and A. C. Carvalho, “Explainable Machine Learning for Breast Cancer Diagnosis”.
- [70] D. Dua and C. Graff, “UCI machine learning repository,” 2017. [Online]. Available: <http://archive.ics.uci.edu/ml>.
- [71] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” *Journal of Machine Learning Research*, vol. 13, 2012.
- [72] F. Doshi-Velez and B. Kim, “Towards A Rigorous Science of Interpretable Machine Learning”.

## Image sources

- [1]. <https://www.intel.it/content/www/it/it/artificial-intelligence/posts/difference-between-ai-machine-learning-deep-learning.html>
- [2]. <https://starship-knowledge.com/supervised-vs-unsupervised-vs-reinforcement>
- [3]. <https://ipullrank.com/resources/guides-ebooks/machine-learning-guide/chapter-2>  
(modified)
- [4]. <https://ipullrank.com/resources/guides-ebooks/machine-learning-guide/chapter-2>  
(modified)
- [5]. <https://ipullrank.com/resources/guides-ebooks/machine-learning-guide/chapter-2>  
(modified)
- [6]. [https://www.researchgate.net/figure/K-means-clustering-algorithm-An-example-2-cluster-run-is-shown-with-the-clusters\\_fig3\\_268880805](https://www.researchgate.net/figure/K-means-clustering-algorithm-An-example-2-cluster-run-is-shown-with-the-clusters_fig3_268880805)
- [7]. [https://ucsdnews.ucsd.edu/pressrelease/why\\_are\\_neuron\\_axons\\_long\\_and\\_spindly](https://ucsdnews.ucsd.edu/pressrelease/why_are_neuron_axons_long_and_spindly)
- [8]. Self-made
- [9]. [http://rasbt.github.io/mlxtend/user\\_guide/general\\_concepts/activation-functions/](http://rasbt.github.io/mlxtend/user_guide/general_concepts/activation-functions/)
- [10]. Self-made
- [11]. Self-made
- [12]. Self-made
- [13]. Self-made
- [14]. Self-made
- [15]. P. Cerveri, Dispense di Neuroengineering, 2019
- [16]. Bahi, Meriem & Batouche, Mohamed. (2018). Deep Learning for Ligand-Based Virtual Screening in Drug Discovery. 1-5. 10.1109/PAIS.2018.8598488.
- [17]. P. Cerveri, Dispense di Neuroengineering, 2019
- [18]. <https://technoelearn.com/convolutional-neural-network-tutorial/> (modified)
- [19]. P. Cerveri, Dispense di Neuroengineering, 2019
- [20]. <https://insightsimaging.springeropen.com/articles/10.1007/s13244-018-0639-9>  
(modified)
- [21]. [https://www.researchgate.net/figure/Feature-maps-of-different-residual-blocks-For-an-intermediate-feature-map-with-C-H-W\\_fig1\\_332696607](https://www.researchgate.net/figure/Feature-maps-of-different-residual-blocks-For-an-intermediate-feature-map-with-C-H-W_fig1_332696607) (modified)
- [22]. P. Cerveri, Dispense di Neuroengineering, 2019

- [23]. [https://computersciencewiki.org/index.php/Max-pooling\\_/Pooling](https://computersciencewiki.org/index.php/Max-pooling_/Pooling)
- [24]. <https://www.analyticsvidhya.com/blog/2018/12/guide-convolutional-neural-network-cnn/>
- [25]. <https://openai.com/blog/deep-double-descent/>
- [26]. <https://www.darpa.mil/program/explainable-artificial-intelligence>
- [27]. <https://trends.google.it/trends/?geo=IT>
- [28]. P. Gohel, P. Singh and M. Mohanty, "Explainable AI: current status and future directions," IEEE Access.
- [29]. A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, 2018.
- [30]. <https://towardsdatascience.com/advanced-topics-in-deep-convolutional-neural-networks-71ef1190522d>
- [31]. N. Tsiknakis, E. Trivizakis, E. E. Vassalou, G. Z. Papadakis, D. A. Spandidos, T. Aristidis, J. Sanchez-Garcia, R. Lopez-Gonzalez, N. Papanikolaou, A. H. Karantanas and K. Maria, "Interpretable artificial intelligence framework for COVID 19 screening on chest X rays," Experimental and therapeutic medicine, vol. 20, 2020.
- [32]. N. Tsiknakis, E. Trivizakis, E. E. Vassalou, G. Z. Papadakis, D. A. Spandidos, T. Aristidis, J. Sanchez-Garcia, R. Lopez-Gonzalez, N. Papanikolaou, A. H. Karantanas and K. Maria, "Interpretable artificial intelligence framework for COVID 19 screening on chest X rays," Experimental and therapeutic medicine, vol. 20, 2020
- [33]. N. Tsiknakis, E. Trivizakis, E. E. Vassalou, G. Z. Papadakis, D. A. Spandidos, T. Aristidis, J. Sanchez-Garcia, R. Lopez-Gonzalez, N. Papanikolaou, A. H. Karantanas and K. Maria, "Interpretable artificial intelligence framework for COVID 19 screening on chest X rays," Experimental and therapeutic medicine, vol. 20, 2020
- [34]. L. Gentiluomo, D. Roessner, D. Augustijn, H. Svilenov, A. Kulakova, S. Mahapatra, G. Winter, W. Streicher, A. Rinnan, G. H. Peters, P. Harris and W. Frieß, "Application of interpretable artificial neural networks to early monoclonal antibodies development," European Journal of Pharmaceutics and Biopharmaceutics, vol. 141, 2019.
- [35]. L. Gentiluomo, D. Roessner, D. Augustijn, H. Svilenov, A. Kulakova, S. Mahapatra, G. Winter, W. Streicher, A. Rinnan, G. H. Peters, P. Harris and W. Frieß, "Application of interpretable artificial neural networks to early monoclonal antibodies development," European Journal of Pharmaceutics and Biopharmaceutics, vol. 141, 2019.
- [36]. <https://www.mdpi.com/2072-4292/12/7/1149/htm>

- [37]. [https://en.wikipedia.org/wiki/Autoencoder#/media/File:Autoencoder\\_schema.png](https://en.wikipedia.org/wiki/Autoencoder#/media/File:Autoencoder_schema.png)
- [38]. J. Seah, J. Tang, A. Kitchen and J. Seah, “Generative Visual Rationale”.
- [39]. T. Brito-Sarracino, M. R. Santos, E. F. Antunes, I. B. A. Santos, J. C. Kasmanas and A. C. Carvalho, “Explainable Machine Learning for Breast Cancer Diagnosis”.
- [40]. T. Brito-Sarracino, M. R. Santos, E. F. Antunes, I. B. A. Santos, J. C. Kasmanas and A. C. Carvalho, “Explainable Machine Learning for Breast Cancer Diagnosis”.
- [41]. T. Brito-Sarracino, M. R. Santos, E. F. Antunes, I. B. A. Santos, J. C. Kasmanas and A. C. Carvalho, “Explainable Machine Learning for Breast Cancer Diagnosis”.