LUISS

Department
of Business and Management

Bachelor's Degree in Management and Computer Science
Course of Artificial Intelligence and Machine Learning

Artificial Intelligence in Medicine:

Real-World Implementations, Risks and Perceptions of Italian Healthcare Professionals on AI's Impact

Supervisor:
Prof. Giuseppe Francesco Italiano

Candidate:
Ivana Jasna Caltagirone
ID No: 248771

Academic Year 2021/2022

Index

List of Figures

List of Tables

Acknowledgements

Vorrei esprimere la mia gratitudine al Prof. Giuseppe Italiano per la sua disponibilità e l'aiuto fornitomi nella stesura della tesi. Questo percorso di studi triennale, per il quale mi sono sentita poco adatta innumerevoli volte e che all'inizio mi sembrava uno scoglio insormontabile, mi ha dato la capacità di acquisire maggiore sicurezza in me stessa e mi ha spinto a continuare a mettermi alla prova. Mi ha insegnato che nessun traguardo è irraggiungibile, e questo, per me, è solo il primo di una serie di traguardi che sogno di raggiungere. La fine di questo percorso genera in me felicità e soddisfazione, ma allo stesso tempo una sorta di malinconia che sorge spontanea ogni qual volta si chiude un capitolo della propria vita. A far parte di questo capitolo ci sono state tante persone a cui devo i miei più sinceri e calorosi ringraziamenti: ai i miei cari amici del corso di Management and Computer Science; a Valeria, che è diventata come una sorella per me in questi tre anni, un'amica su cui posso contare sempre e la cui allegria mi ha accompagnata sin dai primi giorni di università; a Marta, che nonostante la distanza resta sempre la mia ancora di salvezza nei momenti più difficili; alle mie migliori amiche del liceo, il nostro legame è rimasto indissolubile nonostante i percorsi e le strade diverse intraprese; a Marco, che mi ha sostenuta durante tutto l'ultimo anno accademico e la cui impronta è stata fondamentale per poter sviluppare l'idea per questa tesi di laurea. Infine, voglio ringraziare la mia famiglia con tutto il cuore: mia sorella, che mi vuole bene come poche persone al mondo; mio padre, che è sempre stato al mio fianco nei momenti in cui non mi sentivo abbastanza e che è sempre capace di capirmi a pieno, e infine mia madre, il medico più buono, empatico e disponibile che conosco, sono grata per tutti i sacrifici che fai per me ogni giorno pur di non farmi mai mancare nulla ed aiutarmi a realizzare i miei sogni: a te dedico questa tesi di laurea.

Grazie per essere parte della mia vita.

Introduction

In recent years, we have seen the steady introduction and use of Artificial Intelligence in every sphere of work, business, and human life. In some circumstances, it is so entrenched into our daily lives that we are unconscious of its presence and influence.

Progress has accelerated at an incredible rate in the last two decades, allowing previously unimaginable ambitions to be realized. AI now pervades every aspect of human existence: our phone recognizes our preferences and tastes, unlocks using face recognition, suggests which route to take, and tells us in how long we will get from one location to another. Machine Learning techniques have also made significant developments in several industries, including medicine, where it is seen as the great hope of the twenty-first century for enhancing humanity's life prospects. Artificial Intelligence has already brought significant changes in the healthcare system and is expected to bring even more in the coming years: it will provide doctors with increasingly safe and efficient support in collecting and organizing clinical data, facilitating early diagnoses, finding better treatments for patients and many more.

The purpose of this work is to demonstrate the advantages of applying Artificial Intelligence and Machine Learning to medicine, as well as to examine the possibility of implementing such technologies in Italian facilities today or within the next 5-10 years. In the following pages, life-changing evidence as well as significant findings will be introduced through case studies that demonstrate how AI can help physicians' work: the first testimony of this thesis shows how AI can be used to diagnose prostate cancer; in the second case study we'll look at how a machine can make an early diagnosis of COVID-19 by studying sounds; the final section demonstrates the use of AI in the diagnosis of breast cancer.

Along with the benefits, the digital revolution and the entrance of Artificial Intelligence into our routine have raised plenty of questions, worries and ethical concerns, as well as the need to update the legal framework to incorporate these new technologies. Modern algorithms function as "black boxes", which means their operation is not generally understandable "as it is", leading to uncertainty on the way they work and on the liability in the event of machine failures. To solve these complicated legal issues, the creation of a new legislation on Artificial Intelligence, able to fill the gaps left by the current one, is necessary. Moreover, while AI can make a great contribution to medical activity, it is crucial not to overlook the physician’s role, who must constantly review ML accuracy and clinical significance for the workflow to function well.

The final goal of this thesis is to explore the perspectives and awareness among existing and future Italian healthcare professionals – with a major focus on Sicilian and Lombard personnel – on AI and its applications in the medical field. Moreover, the aim of the study is to determine the perceived benefits and risks of AI in medical services, and whether Italy is prepared for this change, considering that it ranks last in terms of digital capabilities according to the DESI (Digital Economy and Society Index). The research, carried out through an online survey which is proposed at the end of this work, investigated the knowledge of the respondents related to the most general definitions and applications of Artificial Intelligence, zooming into the familiarity participants have with AI-based technologies in medicine, their opinion on the most useful fields of applications and their major concerns in the implementation of such systems.

The study was conducted to validate or contradict previously reflected hypotheses according to my perceptions on the subject. Indeed, my assumptions were that the Italian healthcare staff lacked the necessary competencies and familiarity with AI, being also concerned about its adoption, the prospect of losing their job, and the legal responsibilities in the event of machine errors. I also assumed that most respondents have never utilized AI-based medical systems in the facility where they work, and that they were unaware of present applications of AI in medicine today, along with their benefits, demonstrated in Chapter 2. Additionally, I theorized that the participants were very concerned about the misuse of their personal data. Because of all these risks, professionals would be hesitant to change the current healthcare system, hoping that AI would never be introduced, or that the process would take 15 years or longer.

Finally, as showed in Chapter 3, the study yielded noteworthy results and differences between the two regions, allowing for the drawing of conclusions that could lead to a genuine shift in the Italian mode of operation from many perspectives: legal, medical and educational.

Chapter 1

Introduction to Artificial Intelligence

1.1 Historical Background

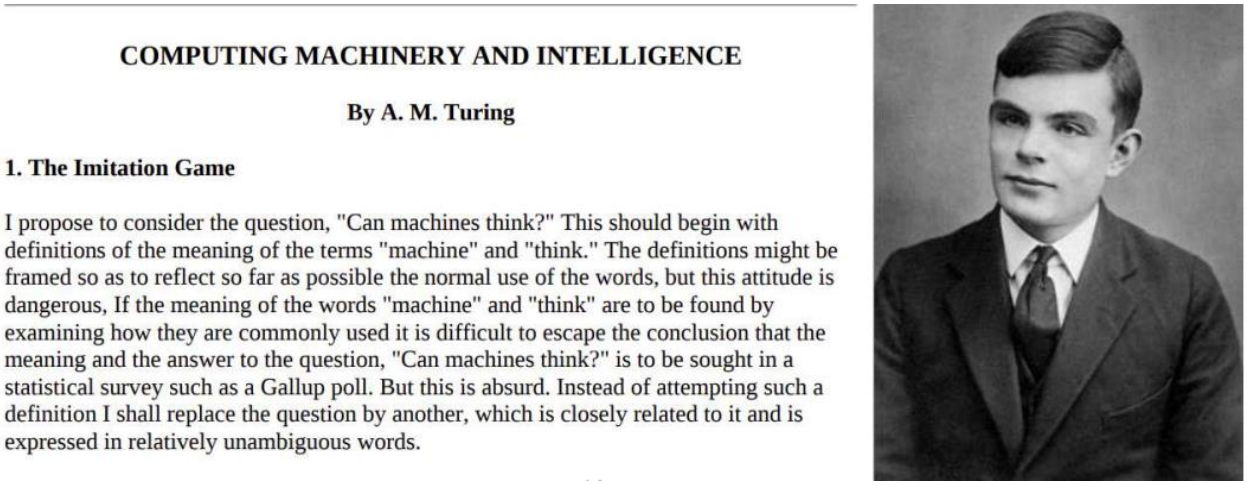
Starting from the first century B.C, humans have been fascinated by the possibility of creating machines that could simulate the human brain. In the modern era, the first to ever lay the foundations for computers up to the present day was Alan Turing, with his article "*On Computable Numbers, With An Application To The Entscheidungsproblem",* published in 1936 and which defined key concepts such as calculability, computability and Turing machine.

Together with the first discoveries related to calculus and computation, many important scientific inventions related to neurology and cybernetic theories led to the question: "Can an electronic brain be built?". It is from this impulse that the idea of a thinking machine originated. Hence, in 1943, McCulloch and Pitts created what can be considered as the first work ever related to Artificial Intelligence: a system that used artificial neurons with a state of "on" and "off", and a transition to "on" thanks to a given number of stimuli that came from a sufficient amount of surrounding neurons. Seven years later, in 1950, Marvin Misky and Dean Edmons created the first artificial neural network, known as SNARC.

In the same year, Alan Turing [Figure 1] published the paper "Computing machinery and intelligence", introducing the concept of Turing Test to the general public. The fundamental question Turing proposed was *"Can machines think?"* and in order to provide and answer to his dilemma, the mathematician explained a criterion in order to determine if a machine could exhibit an intelligent behavior through an experiment known as the **"Imitation Game"**. With the term "intelligent machine", Turing meant a machine able to think, concatenate ideas and express them with a real meaning, but rather than trying to determine if the machine can think, Turing suggested we should wonder if the machine is able to win a game.

In such game, which is played by three people, a man (A), a woman (B), and an interrogator who can be of either sex (C), the interrogator stays in a room separate from the other two and can communicate with the players by writing notes or through any other form of communication that does not reveal any information about the gender of A and B. The purpose of the game from the interrogator's perspective is to determine which of the two is the man and which is the woman by

asking them questions. Turing then proposed a variation of the game that involved a computer: *"We now ask the question, what will happen when a machine takes the part of A in this game? Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?"* These questions replace our original, "Can machines think?". If the percentage of times in which C guesses who the man and the woman are is similar before and after the substitution with the machine, then the machine itself should be considered intelligent, since – in this situation – it would be indistinguishable from a human being.



COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. The Imitation Game

I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous, If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

Fig. 1: *Alan Turing*

In 1956, In New Hampshire, at Dartmouth College, a conference dedicated to the development of intelligent systems was held and attended by some of the leading figures in the emerging field of computation: John McCarthy, Marvin Minsky, Claude Shannon and Nathaniel Rochester. On John McCarthy's initiative, a team of ten people were to create in two months a machine capable of simulating every aspect of human learning and intelligence. In the same conference, McCarthy introduced the expression **"Artificial Intelligence"**, which marked, in an indelible way, the actual birth of this discipline, giving it its own nature that we still distinguish nowadays.

Among the various aspirations on the side of the researchers, the main one was creating machines capable of exhibiting human-like reasoning abilities. For instance, Herbert Simon, in 1957 estimated that within ten years there would be machines capable of competing with chess champions (prediction that will come true, but after forty years). This aspiration had to collide with some difficulties: first of all, the absolute lack of knowledge about the domains treated by the machines, as their reasoning ability was limited to a mere syntactic manipulation, and other uncertainties related to the logic of computers. In those days, the first programming language was born: its name was **Lisp** (List Processor), created by McCarthy in 1958.

In the years following the conference, the development of AI increased further: in 1966, the first chatbot known as ELIZA was born and developed by Joseph Weizenbaum at the Massachusetts Institute of Technology (MIT). ELIZA, communicating via text in human language rather than computer code, was an early example of natural language processing. ELIZA was the ancestor of today's chatbots, such as Alexa and Siri, which can now communicate with speech in addition to text. Thanks to the progresses reached, the researchers' predictions about the future of AI were optimistic, with computers performing an increasing amount of tasks, from speaking English to solving algebraic equations, but this raised an additional problem: the inability to process the amount of data required for the application to be successful. This led from the first excitement related to AI to an era full of uncertainties and struggles that lasted until the mid-1990. In those years, we faced a progressive increase in computing power, and in 1997, a remarkable event occurred: the Deep Blue computer software managed to beat the world chess champion Garry Kasparov. Another point in favor of AI over experts was scored in 2016 when Deep Mind's AlphaGo defeated 18-time world champion Lee Sedol.

In the last twenty years, progress has advanced at a frightening pace, enabling unimagined goals to be achieved. AI now pervades every aspect of our daily lives: our cell phone knows our preferences and tastes, unlocks via facial recognition, indicates which route to use, tells us how long it will take to get from one place to another. These series of events led to the creation of Machine Learning, deep learning, predictive analytics, and, most recently, prescriptive analytics. They also gave rise to an entirely new field of study, namely data science.

1.2 What is Artificial Intelligence

Artificial Intelligence is a field of computer science that is concerned with designing intelligent computer systems, that is, systems that exhibit the characteristics we associate to intelligent human behavior. These intelligent computer systems should be able to solve problems autonomously and without humans’ aid, making use of logic, if-then rules, decision trees and Machine Learning. An intelligent system is expected to display the following capabilities:

- **Reasoning and problem solving:** Initially, the researchers concentrated on constructing algorithms that completely imitated step-by-step human reasoning, beginning with a symbolic representation of the state of the world and searching for actions that achieved the desired aim. Later, such strategies expanded to account for more complicated features such as information uncertainty or incompleteness, probability, statistics, and economics.

- **Learning:** An AI should learn by adding new information to what it already knows. The breadth and speed with which an AI system absorbs knowledge determines its worth.
- **Knowledge Representation:** The quality of the results is determined by the intelligent system's ability to represent its knowledge. Knowledge and its representation are especially crucial for that category of intelligent systems that base their behavior on an extensive, explicit knowledge depiction of their operating environment.
- **Planning:** Intelligent systems must be able to establish and pursue goals in order to forecast and depict future states of the world, as well as make decisions to reach those states while maximizing the expected value of actions. In traditional planning challenges, an intelligent system can assume that it is the only entity functioning in the environment and that it can be certain of the implications of any action made. If an intelligent system is not the sole player in the environment or if the environment is not deterministic, it must constantly assess the results of its activities and update its future forecasts and plans.
- **Natural Language Processing:** The capacity to process natural language provides intelligent systems with the possibility to read and understand language used by humans. This capability is essential in all applications of artificial intelligence that require searching for information, answering questions, translating or analyzing text. The main difficulty in this process is the inherent ambiguity that characterizes natural languages, so solutions require significant knowledge of the world and considerable expertise in managing it.
- **Motion and manipulation:** Artificial intelligence is commonly employed in robotics. Robots are intelligent systems for all tasks that require cognitive level capabilities for manipulating or moving objects and for locomotion, with the sub-problems of localization (determining one's position and that of other entities in the space), map construction (learning the characteristics of the surrounding space), movement planning and execution.

1.3 Fields of application of Artificial Intelligence

Artificial Intelligence has been applied in a wide range of sectors and applications, including medicine, the stock market, robotics, law, scientific research and even toys. In certain applications, Artificial Intelligence has grown so ingrained in society or industry that it is no longer recognized as such. Few of the many AI applications are:

- **Medicine:** The use of Artificial Intelligence in medicine simplifies the lives of healthcare workers by allowing them to accelerate the process of diagnosis and scientific research, thereby saving a growing number of lives. The subsequent chapters will explain and illustrate the spread and application of AI in support of clinical staff activity.

- **Everyday life:** many of the amenities of modern daily life are provided by the development of AI applications, influencing the way we live, work and have fun. Voice assistants such as Siri, Alexa and Cortana help us finding information, directions, adding events to our calendar, sending messages, etc. They learn automatically based on the user's input and interactions, allowing them to become "smarter", predicting and understanding our questions. Even music apps such as Spotify make use of Artificial Intelligence for suggesting the songs we may like based on what we are used to listen to. Using voice recognition, Spotify's algorithm will also determine the listener's emotional state, gender, age or accent, attributes that it will subsequently use to refine the selection of recommended content.
- **Automotive Industry:** Artificial Intelligence will transform cars in the near future as many companies such as Hyundai, Yamaha, Volkswagen and others are working on different algorithms and have developed their solutions for self-driving cars. AI also provides ways for drivers or passengers to alleviate stress, discomfort, anxiety, drowsiness, temperature maintenance, humidity, climate, and display enhancement.
- **Law:** Artificial Intelligence has also many applications in the judicial field through document review systems, contract analysis, predictive analysis of the results of legal proceedings, chatbot for legal assistance and more. It is hence useful for facilitating the most mechanical and repetitive tasks and thus allowing a streamlining of the workload borne by professionals, who will have more time to focus on more value-added tasks.
- **Finance:** The financial sector was among the earliest adopters of AI for financial technology (or fintech), and its popularity among financial institutions is steadily growing. According to recent research published by Gartner, Artificial Intelligence and Machine Learning tools rank as the best "breakthrough" technology in financial services.
- **Advertising:** The use of big data and Machine Learning software allows for a more precise identification of the targets that the advertiser intends to address. The segmentation is made on the basis of users' online behavior, geographical area of origin and/or their interests.

Related to the discipline of Artificial Intelligence are two areas of study: Machine Learning and Deep Learning.

1.4 Machine Learning

The term Machine Learning was coined by Arthur Samuel in 1959, who defined it as the *"field of study that gives computers the ability to learn without being explicitly programmed".*

It is a branch of Artificial Intelligence [Figure 2] that brings together methods such as pattern recognition, artificial neural networks, adaptive filtering, dynamical systems theory, image processing, data mining etc., and uses statistical methods so to improve the performance of an algorithm in identifying patterns in data. Therefore, algorithms are the pillars of Machine Learning: an algorithm corresponds to an ordered and finite sequence of elementary steps that lead to a well determined result in a finite time.

By training algorithms from a sample set of data, called training set, the machine can learn, make predictions and correct itself, improving constantly. Possible applications include email filtering to avoid spam, optical character recognition, search engines and computer vision.

The main goal of Machine Learning is for a machine to be able to accurately complete novel tasks which it has never faced, after gaining experience on a set of learning data. Machine learning tasks are typically classified into three categories, selected depending on the nature of the data available.

The three categories are:

1. Supervised Learning;
2. Unsupervised Learning;
3. Reinforcement Learning.

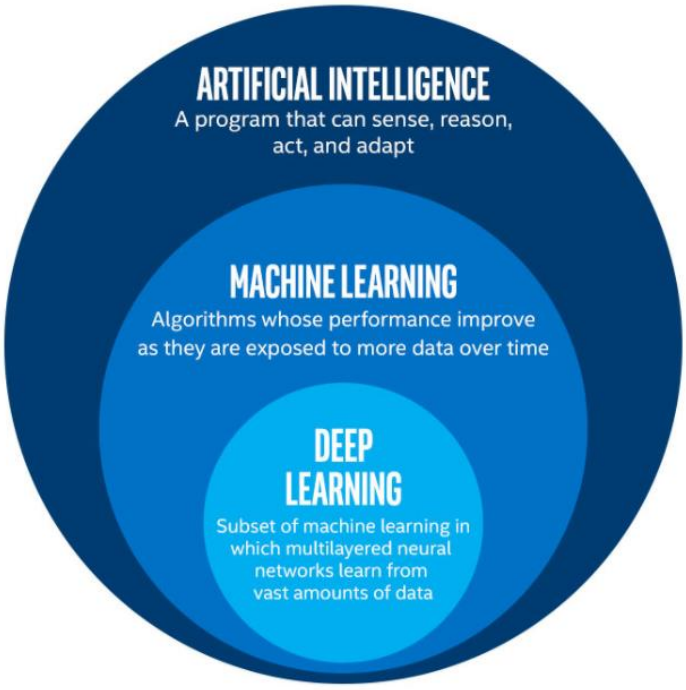


Fig. 2: *AI, ML and DL.*

1.4.1 Supervised Learning

The main purpose of supervised learning is to train a model starting from labeled or pre-categorized training data, allowing us to make predictions about future and never-before-seen data. The dataset we feed to the algorithm is composed by a set of training examples (the input data), and the desired output or the desired solutions, called labels, which are already known. The main goal is to extract a general rule that associates the input to the correct output. Hence, supervised learning techniques aim to instruct a computer system so that it can automatically make predictions about the output values of a system with respect to an input that is initially provided to it.

Such algorithms can be used in a wide variety of fields: examples include the **medical field,** in which particular conditions can be predicted based on the experience of past biometric data; speech identification that improves based on past audio listening; handwriting identification that improves on observations of examples submitted by the user.

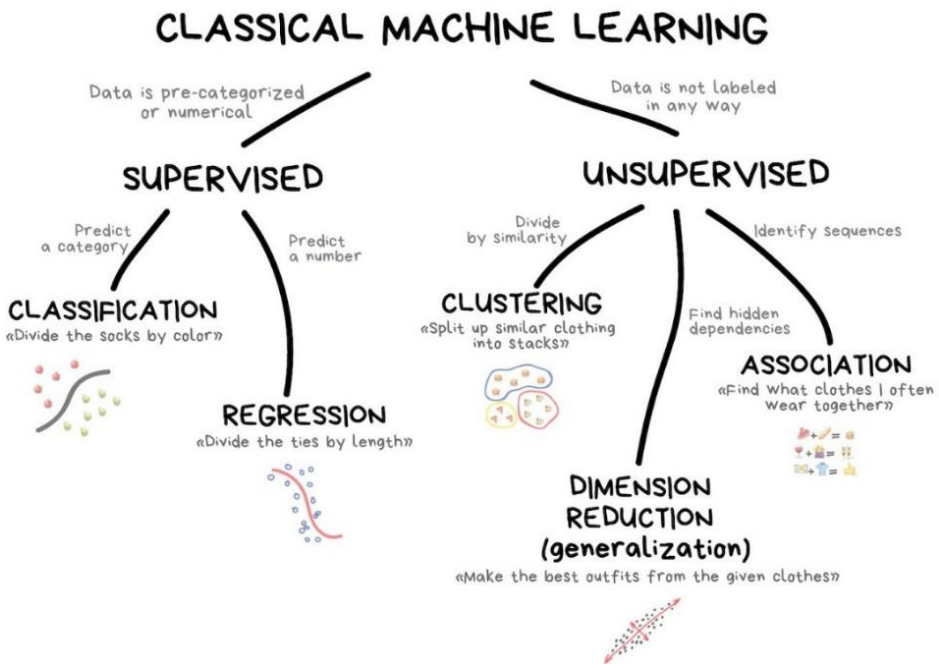


Fig. 3: *Supervised vs Unsupervised Learning*

Fundamental to supervised learning is the **training set**, as mentioned before, in which a response (output) is automatically generated in relation to the input. If the response corresponds to the correct one, then there is a reinforcement of the techniques that led to the input; contrary, that is if the answer is not that one expected, the algorithm is modified in a way to optimize the result and diminish the deviation from the correct solution. Subsequently we pass to a phase of test in which the ability of generalization of the algorithm is verified (**test set**). Hence, the test set is used to

determine the accuracy of the model. The test set needs to be large enough to yield statistically meaningful results, and it needs to represent the data set as a whole. Supervised Learning techniques include **Classification** and **Regression**: when the desired output is a class, a category, or a finite set of values (such as true/false, 1 or 0), then we are dealing with a Classification problem. When the output is a number (such as the expected price of wine in the future), then we are dealing with a Regression problem.

In **Classification algorithms**, the main goal is understanding which category a new observation belongs to, and hence, "classify" it on the basis of the training set. An example of a classification algorithm can be the Email Spam Detector algorithm: it is trained with many example emails along with their class (spam or non-spam), and it learns how to classify the upcoming emails.

In **Regression algorithms**, the relationship between a dependent variable (the outcome) and one or more independent variables is estimated, and the output is a real (continuous) value. The goal is to establish if there exists a sort of correlation between the independent and the dependent variables. Regression can be used to estimate the value of a currency's exchange rate in the future based on its recent values, or to determine the price of wine given some independent variables such as the wine's age, weather, and harvest rain. There are various classification and regression algorithms available, with the most notable ones given below.

Linear Regression

Simple Linear Regression predicts an outcome variable (dependent variable) given one independent variable (one predictor) [Figure 4], while Multiple Linear Regression makes use of more than one independent variable. With linear regression, we fit a line to the data using "least squares": in other words, we find the line that minimizes the sum of the squares of the **residuals**, which are the differences between our predictions and the real outcome values.

The Simple Linear Regression model is given by:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Where:

- i varies between the observations, $i = 1, \ldots, n$;
- Y_i is the dependent variable, or outcome for the i-th observation;
- X_i is the independent variable or regressor;
- β_0 is the intercept coefficient;
- β_1 is the regression coefficient for the independent variable;

- ε_i is the statistic error for the i-th observation.

Linear regression makes several assumptions about the data at hand:

1. **Linearity:** There exist a linear relationship between dependent and independent variables;
2. **Independence of the errors:** there is no relationship between the residuals and the outcome variable: errors are not influencing each other;
3. **Multivariate normality:** any linear combination between variables is normally distributed;
4. **Homoscedasticity:** The error term is the same across all values of the independent variables;
5. **No Multicollinearity:** Multiple linear regression assumes that the independent variables are not highly correlated with each other.

In order to determine if the linear regression model is reliable, there are different metrics based on the residuals that help measuring the size of the error that our prediction generates. Two of the most useful are:

- **RMSE (Root Mean Squared Error):** an absolute error measure in which deviations are squared to prevent positive and negative values from cancelling each other out. With this measure, larger value errors are also amplified, a feature that can facilitate the elimination of methods with the most significant inaccuracies. Technically, RMSE provides the standard deviation of the residuals, which explain how far the data points are from the regression line. The value can lay between 0 and 1: the closer to 1, the most accurate the model.
- **R^2:** an indicator of quality that measures the proportion of the variation in the dependent variable that is predictable from the independent variables, comparing it to a "base case" model, the value of the mean.

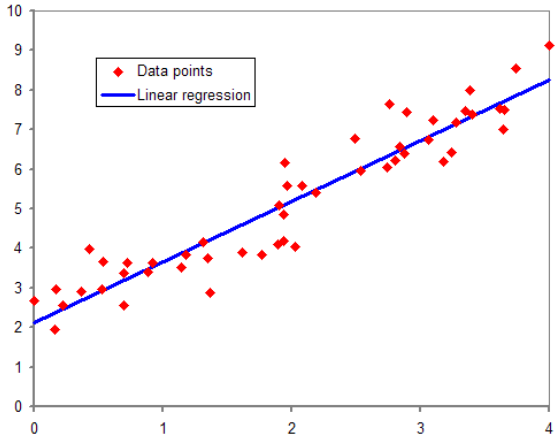


Fig. 4: *Linear Regression*

It captures the strength of the relationship between our model and the dependent variable on a range from 0 to 1: the closer to 1, the higher the accuracy.

Logistic Regression

Logistic regression handles classification problems to predict the class or category of individuals based on one or multiple predictor variables. It is a nonlinear regression model used when the outcome is binary, which means that it can assume two modes: an example is "gender", which can only take on two values: male and female. Hence, logistic regression measures the probability of occurrence of two different outcomes, and sets a threshold to classify each observation into one of the two classes. The curve goes from 0 to 1, which means that the outcome is not the class itself, but the **probability of belonging to one of the two categories** [Figure 5].

Differently from linear regression, which fits a line to the data, logistic regression fits an "S" shaped "logistic function". For this reason, logistic regression does not have the same concept of "residual" as linear regression, so it cannot use "least squares" or calculate R^2. Instead, logistic regression builds a **confusion matrix** to summarize the classifications and calculates the number of correct and incorrect classifications using false negatives, false positives, true negatives and true positives. Then, it measures the **Sensitivity** and **Specificity** to evaluate the model with the given threshold for the probability. Sensitivity calculates the percentage of true positives, while Specificity captures the percentage of true negatives.

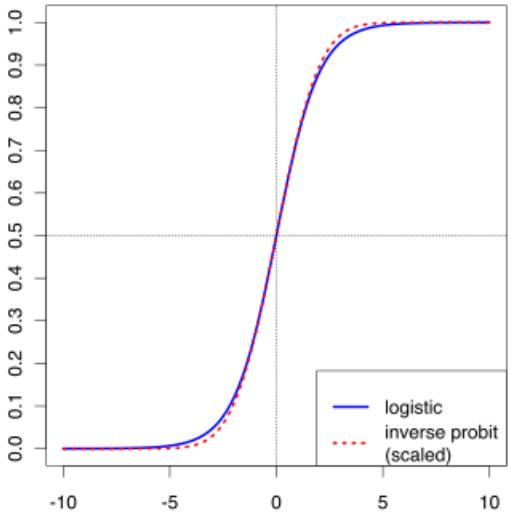


Fig. 5: *Logistic Regression*

A helpful technique to measure the accuracy of our model and capture all thresholds simultaneously is the **Receiver Operator Characteristic (ROC)** curve, which provides on the y-axis the Sensitivity (True Positive Rate) and on the x-axis the False Positive Rate (1 – Specificity). An

additional useful method is the **Area Under the Curve (AUC):** the bigger the area under the curve, the more accurate the model.

K-nearest neighbors (K-NN)

K-Nearest Neighbors is a supervised regression and classification algorithm that works by leveraging similarities among examples, so to assign them a specific category. The most common way in which the algorithm measures similarities between different data points is through the "**Euclidean distance".**

The Euclidean distance ***D*** between two points ***x*** and ***y*** is given by:

$$D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

In pattern recognition, K-NN classifies objects based on the features of the data points that are the closest (measuring the Euclidean distance) to the one being considered. The input is the K closest training example in the feature space, and the output depends on whether K-NN is used for classification or regression.

- In **K-NN classification**, the outcome is a class membership. An object is classified by a plurality vote of its neighbors, and it is then assigned to the most common class among its K-Nearest Neighbors (k is a positive, odd integer to avoid parity and it is typically small).
- In **K-NN regression**, the outcome is the value of the object, which is given by the average of the values of the neighbors.

Deciding the right parameter K is fundamental for a good prediction of the category or of the number. When K is too small or equal to 1, the object will be assigned to the class of the closest neighbor, and thus the classification might be incorrect. On the contrary, for larger values of K, our predictions become more stable due to majority voting and thus more likely to be accurate up to a certain threshold. Eventually, we begin to witness an increasing number of errors: at that point, we know we have pushed our value K too far.

Figure 6 shows an example of classification through K-NN. The data point to classify is given by the yellow question mark. The two categories the object might belong to are:

1. Class A: red stars.

2. Class B: green triangles.

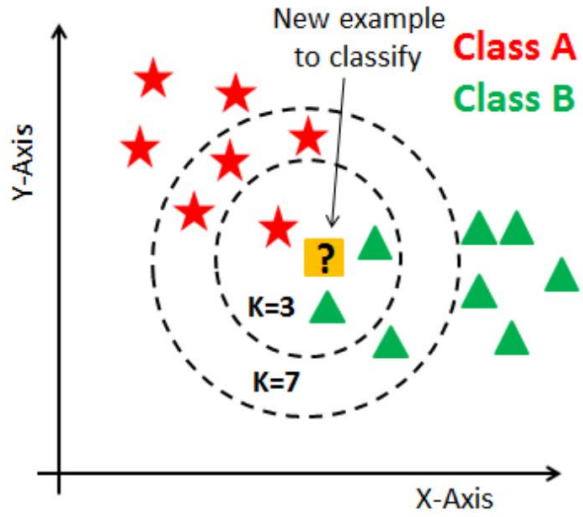


Fig. 6: *K-NN*

The algorithm follows four steps:

1. Choose the number of K neighbors (in the first case, K=3, in the second case, K=7);

2. Take the K nearest neighbors of the new data point according to the Euclidean distance;

3. Among these K nearest neighbors, count the number of data points in each category;

4. Assign the new data point to the class with the majority of nearest neighbors.

In the first case, with K=3, the new data point will belong to Class B, because there are two green triangles and only one red star. In the second case, choosing K=7, the new data point will belong to Class A, because it counts four red stars and only three green triangles.

Such algorithm shows many **advantages**:

1. It is more interpretable than other Machine Learning techniques such as logistic regression;

2. There is no explicit training required, because the model uses all the training points for classification and stores the results in memory;

3. It is more flexible on non-linear data, because it does not assume linearity on the training set.

It also presents some **disadvantages**:

1. Leads to high memory usage and is computationally expensive in case of a large data set;

2. It is sensitive to missing values, noisy data and outliers.

Support Vector Machine (SVM)

SVM is a supervised regression and classification algorithm in which the main goal is to assign a new data point to a class by finding a separation line that maximizes the margin between two possible classes. The margin refers to the minimum distance between the points in the classes and the separation line itself.

Thus, a SVM model represents examples as points in space, mapped such that points belonging to the two classes are clearly separated by as large a gap as possible. This is achieved by defining a separation line that achieves the greatest margin between the two closest examples from each class. New points are then mapped into the same space, and the prediction on the category to which they belong is determined by the side in which they fall. This is done by using a small part of the training set, the so-called **"support vectors"**: these are the values of one class closest to the separation line, hence those that can be classified with the most difficulty. Looking at Figure 7, the star and the triangles are the support vectors that lay on the margin: once such vectors are measured, all the other data points will not influence the classification.

In addition to linear classification, it is possible to make use of SVM to effectively perform nonlinear classification (for non-linearly separable data) using the kernel method by implicitly mapping their inputs into a multi-dimensional feature space. SVM works well when there is a clear margin of separation between the categories and it is more effective in high dimensional spaces compared to K-NN, but unfortunately it can become very computationally expensive.

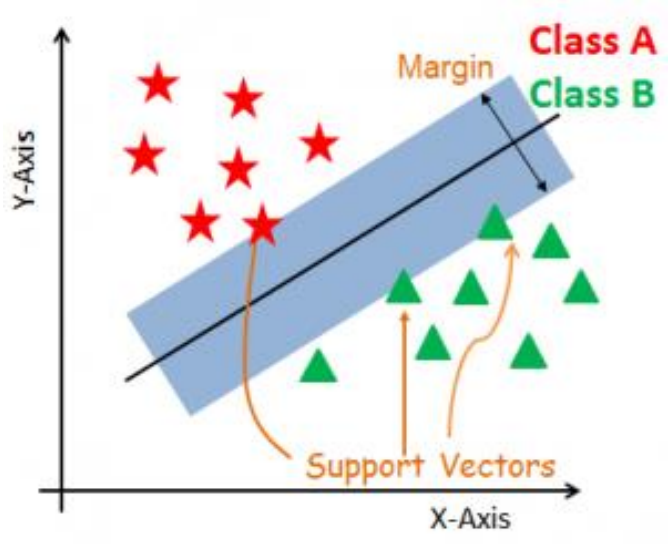


Fig. 7: *SVM*

Classification and Regression Trees (CART)

A Classification and Regression Tree is a predictive algorithm useful for both classification and regression problems. By making use of Decision Trees, the model predicts the value of the outcome based on the values of the predictors. The tree is composed by:

- The **root node** on the very top;
- **Internal nodes** representing the input variables;
- **Edges** which connect the internal nodes and give the possible value for the condition set in the decision nodes (Yes or No, True or False…);
- **Leaf noes** at the end of the tree, which represent the outcome.

Thus, to predict the most frequent outcome of an observation, it is enough to follow the edges from the root to the end. Since each subset of the tree can show up more than a single outcome, the percentage of data in the subset is computed and a threshold must be set to obtain a prediction. CART is one of the best Machine Learning models since it is highly interpretable, not computationally expensive and handles both numerical and categorical data. Unfortunately, decision trees are prone to overfitting and inaccuracy, failing to adapt in predicting future observation reliably, mostly when they create deep, complex trees that cannot generalize well the training data.

In the example below, [Figure 8] we want to predict the class to which a flower having feature *"petal length 4.5 cm"* belongs to. Observing the conditions of the decision nodes, starting from the root down to the leaves, we can clearly see that the most frequent outcome for such observation is "versicolor".

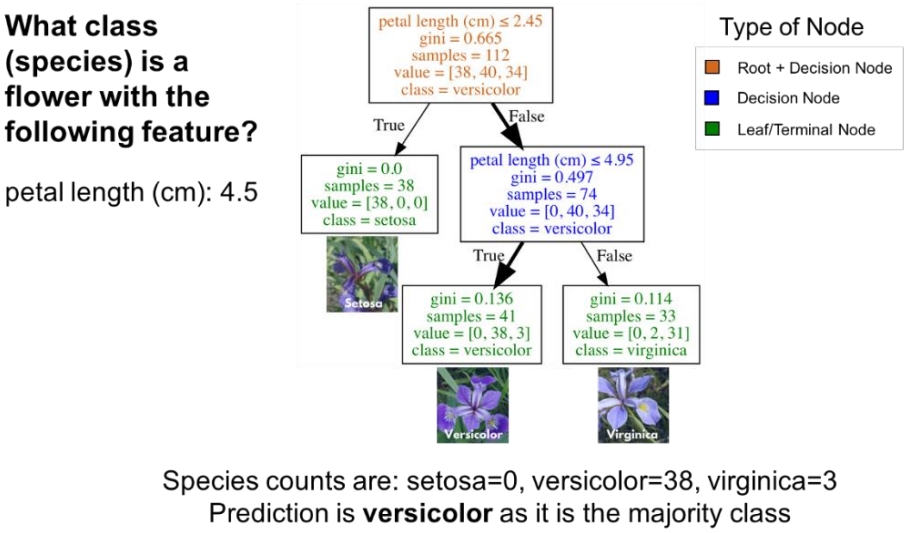


***Fig. 8**: CART*

Similar to CART, Random Forests are algorithms that work by building a large number of Classification and Regression Trees, leading to a vast improvement in accuracy, but also to a more computationally complex and far less interpretable model. To make a prediction for a new observation, each tree "votes" on the outcome, and the result that led to the majority of votes will be considered the final prediction.

1.4.2 Unsupervised Learning

The main purpose of unsupervised learning is to train a model starting from unlabeled data. It consists of providing the algorithm a set of inputs that it will reclassify and reorganize based on common characteristics, carrying out reasoning and predictions about the subsequent ones. Contrary to supervised learning, where the data provided is labeled, the classes are not known a priori but discovered automatically later. A typical example of these algorithms is a search engine: given a set of keywords, the machine can create a list of relevant web pages that are related to the word typed. Hence, unsupervised learning algorithms work by finding similarities or differences in the data they compare, doing so very efficiently in case of numerical inputs, but often struggling with non-numerical data. Essentially, the model works correctly when the data provided has a clearly identifiable grouping. The most common unsupervised learning algorithm is "clustering".

Clustering

Clustering aims to search, select and group homogeneous elements in a data set based on similarities. The algorithm does so by measuring the distances between data points, and therefore the condition of belonging or not to a set depends on how far an element is from the set itself. The main goal is to find the largest number of homogeneous groups that are as different as possible between each other: more specifically, to maximize the variance between the clusters (interclass variance) and to minimize the variance of the observations that belong to the same group (intraclass variance). Two techniques, differentiated through the type of algorithm used to divide the space, are the most common:

- **K-means clustering:** aims at minimizing the within-cluster variance, which is the sum of squared (Euclidean) distances between each data point of the cluster and the closest cluster center (or centroid) of their containing cluster. It works by selecting the number of clusters K to be created and choosing randomly K objects from the data set as initial cluster centers. Then, it assigns each observation to their closest centroid, and it recomputes and re-assigns each data point to the closest cluster center iteratively, in order to minimize the total within sum of squares.

- **Hierarchical clustering:** aims at building a hierarchy of clusters based on partitions characterized by an increasing or decreasing number of groups. At the start, each element belongs to a standalone cluster which is then gradually merged two by two with the closest data points into a single cluster. The clusters can be visualized by means of a tree representation, the so-called "dendrogram", in which the steps of grouping are represented [Figure 9].

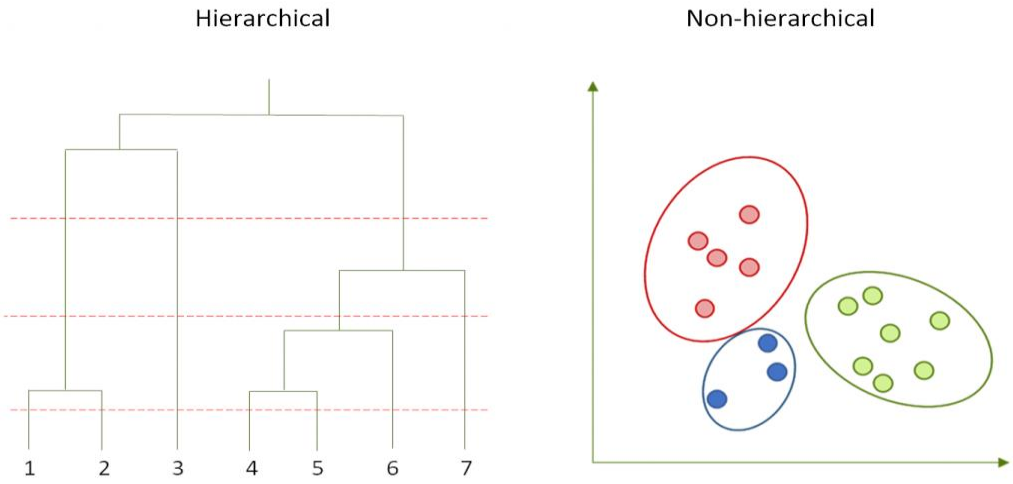


***Fig. 9**: Hierarchical vs Non-hierarchical clustering (K-means)*

1.4.3 Reinforcement Learning

Reinforcement Learning is a Machine Learning technique that aims at realizing algorithms in which the software agents can act in an environment in a way that maximizes their reward. In such algorithms, the agent performs an action and receives a feedback from the environment that will influence all the agent's actions in the future. The feedback, or reward, can either be positive (which will teach the agent to keep performing that action) or negative (which will teach the agent to stop performing that action). Although the activity of the agent may be useful in the present state, it can be counterproductive in future states. On the other hand, actions that seem useless today might be very useful in the future.

The environment in which the agent performs actions is **stochastic**: this means that given a state, all the agent's future actions will depend on the present state and not on the sequence of events that preceded it. Furthermore, we cannot be sure that the action the agent is going to perform will be positive, because the process is **non-deterministic**: it is always influenced by a probability given by the environment response.

Many people refer to reinforcement learning as "**Q-Learning**". Such method assigns a value, or a weight, to each action the agent is going to perform: the environment is explored in a stochastic way and the reward is given based on the choice of the action which generates maximum utility.

Another important reinforcement learning algorithm is "**Temporal differences learning"**: the future rewards will be predicted given the current knowledge, and the algorithm modifies as time changes.

1.5 Deep Learning

Deep Learning is the area of Machine Learning inspired by the structure and functioning of the human brain: it is based on Artificial Neural Networks (ANN) trying to mimic human neurons. ANN influence each other through synaptic connections using multiple layers that progressively, starting from an input layer with a raw input value, extract computations in hidden layers and output features in the output layer. The adjective "deep" refers to the quantity of layers belonging to the ANN, which can be infinitely many. For instance, in image recognition, the raw input value might be a matrix of pixels, the first hidden layer might abstract the pixels and encode the edges; the second hidden layer might encode the arrangements of edges, the third layer might encode mouth and nose, and the output layer will recognize the image of a face.

Learning can be both supervised and unsupervised: such systems learn by considering examples. For instance, in image recognition they might learn to identify an image containing dogs by analyzing similar images that have been already recognized and classified as "dogs" or "no dogs".

ANN, like the real human brain, are composed by artificial neurons that are connected between each other (just like our synapses) transmitting signals to one another. Artificial neurons have the following characteristics:

- They are connected;
- Each connection has a weight that varies the learning process, increasing or decreasing the strength of the connection;
- The input received by a neuron is given by the product between the raw input value coming from such neuron and the weight value of that connection;
- The total raw input value in the input layer is given by the summation of the products of all inputs and their weights;
- Neurons use an *activation function* to evaluate the previous result, transforming it in a non-linear way;

- The information is finally sent to the output layers.

There exist different activation functions: the most used ones are the sigmoid and the hyperbolic tangent functions. Furthermore, it is important to standardize or normalize the input values so to work well with the data. Unfortunately, ANN might make unprecise predictions, generating an output that is different from the real value.

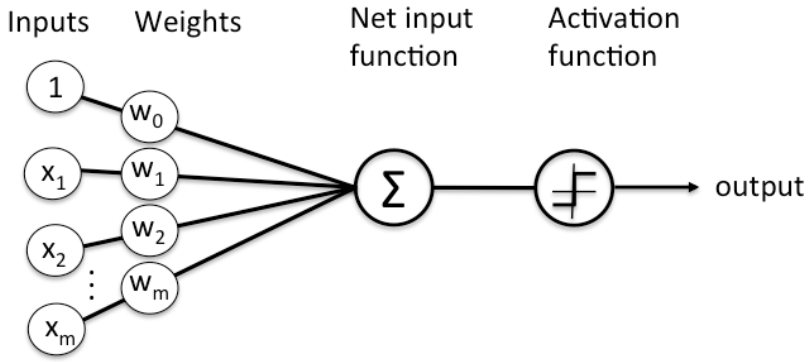


Fig. 10: *ANN in detail*

Thus, the algorithm must adjust the weights of the connections so to reach the least error possible, measured through a *cost function*. Dealing with many data points, our ANN will need to make comparisons between the different errors generated by the different prediction values. So can be done efficiently by training the ANN with the **Stochastic Gradient Descent,** an iterative method for the optimization of differentiable functions that looks at one sample per step. It requires the cost function to be convex, so to find a global minimum to minimize the error. Differently from the simplified version **Gradient Descent**, where the cost function is computed for *all rows* of the dataset to adjust the weights and repeat the operation, Stochastic Gradient Descent updates the weights after calculating the cost function for *each row* of the dataset, taken randomly. Again, the dataset is retraced in epochs. This approach guarantees a greater fluctuation of the cost function and avoids, therefore, to remain blocked in a local minimum. Moreover, although one might be led to think otherwise, Stochastic Gradient Descent is also faster: it does not need to store the whole dataset before calculating the gradient, because such step is done record by record. Furthermore, due to its randomness, this methodology can lead to different results at each training even starting from the same dataset, as opposed to the simple Gradient Descent which is deterministic and always outputs the same result.

Backpropagation

Stochastic Gradient Descent employs backpropagation in the learning process, so to propagate the error back, calculate the gradient of the cost function with respect to all the weights in the network and adjust them to reach the minimum error possible.

In backpropagation, the learning algorithm follows 5 steps:

1. The weights of the connections are initialized randomly;

2. The total raw input data multiplied by the weights is propagated for then to pass the results through the activation function;

3. The obtained results are compared with the actual output;

4. The error is evaluated to understand the goodness of the weights adopted;

5. The actual backpropagation phase, where weights are adjusted if necessary.

Finally, the steps from 1 to 5 are repeated and the weights are updated after every observation.

After the entire training set has been analyzed, the algorithm of backpropagation can be interrupted when the error becomes sufficiently small. Essentially, to train a neural network means to modify in a recursive way parameters, that is "weights" that are initially assigned in a random way and then adjusted every time.

To sum up, backpropagation involves two phases: one that proceeds forward and one that proceeds backward. The phase that proceeds forward presents an example to the neural network, then compares it with the actual output and calculates the error; the phase that proceeds backward propagates the error back into the network to adjust the weights.

Convolutional Neural Network (CNN)

Convolutional Neural Networks are a fundamental tool in the field of Deep Learning. In particular, they are often used in image recognition, and inspired by the organization of the visual cortex. A CNN is a network made up of multiple stages, where each stage is specialized in different things.

The Convolutional Neural Network consists of an input block, one or more hidden blocks (hidden layers), which perform calculations through activation functions (e.g., RELU) and an output block that performs the actual classification. The difference with respect to classical networks is represented by the presence of convolutional layers, which perform a very important task as they extract, through the use of filters, features of the images whose content is to be analyzed. In trying

to recognize and differentiate a specific feature from another, the CNN takes an input image and outputs the class to which the image belongs to. The process follows 3 steps: Convolution, Pooling and Flattening, becoming then a Fully Connected Neural Network.

Step 1: Convolution

The convolutional layer is the initial part of our network where the "Convolution" happens, from which the CNN takes its name. In convolution, image features are extracted thanks to a specific filter that detects the different properties by looking for similarities in the input, known as "**Feature Detector**". The application of such filter allows for highlighting the details of the images so that it is possible to find their meaning and identify the objects.

In this step, the size of the input is reduced so that it can be processed faster. Some information is being lost, but only those ones that are not important for the model, while the meaningful ones will be preserved in the "**Feature Map**". Practically, as we can observe from Figure 11, an input image is divided in squares and multiplied by a matrix (the feature detector), obtaining a Feature map. Many Feature Maps will be created to generate the first Convolutional Layer, where the so-called **ReLU activation function** will be applied to increase non-linearity, being images highly non-linear.

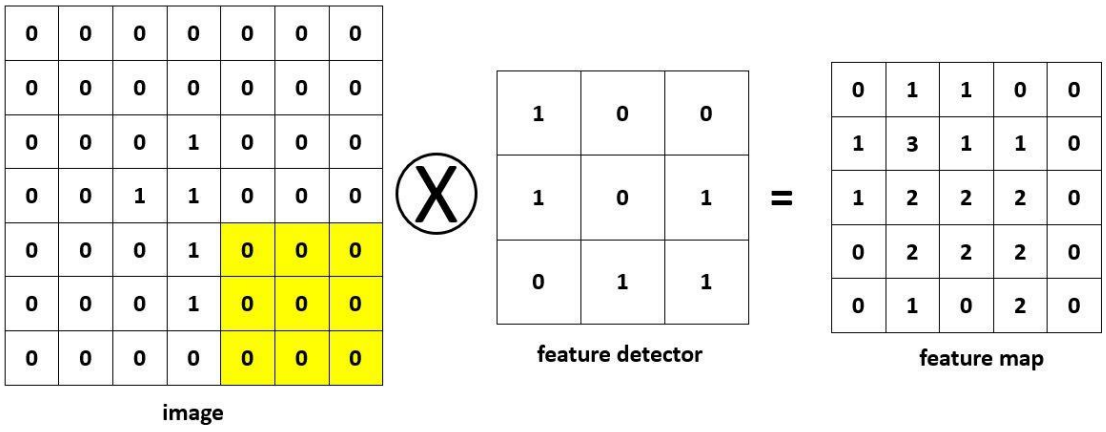


Fig. 11: *Convolution*

Step 2: Pooling

The feature we want to recognize does not always appear in the same position: for this reason, it needs to be rotated and changed so that additional information can be extracted from the input. Such process is called **Max Pooling,** a second class of filter useful for grouping and filtering the pixels into a subset to reduce the size of the image further, while maintaining the original characteristics and represent the closest similarities through the largest numbers on the new matrix. The Feature

Map becomes then, through Max Pooling, a **Pooled Feature Map**. Such process is showed in Figure 12.

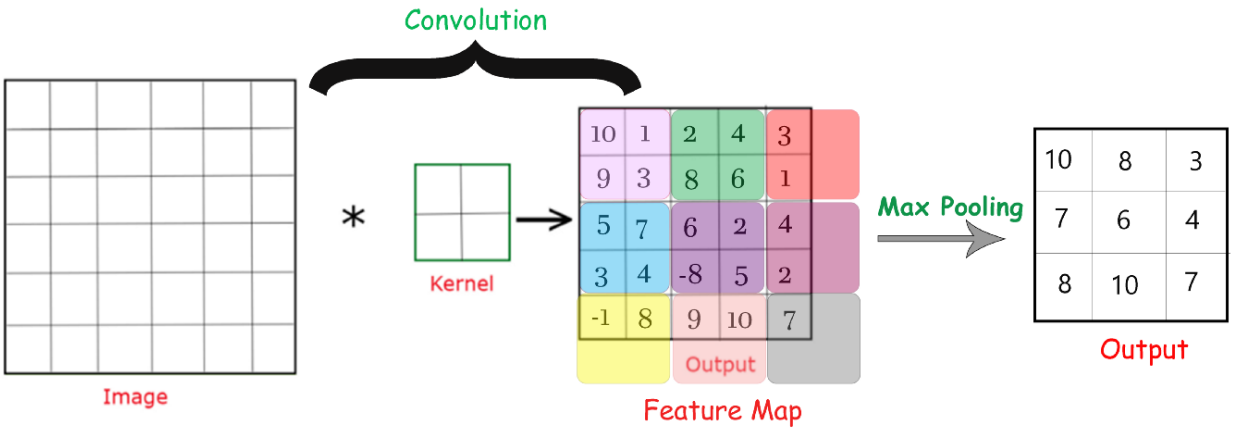


Fig. 12: *Max Pooling*

Step 3: Flattening

Since Artificial Neural Networks cannot work with matrices, the Pooled Feature Map is "flattened" into a column as showed in the image below, to create the input layers for a future ANN.

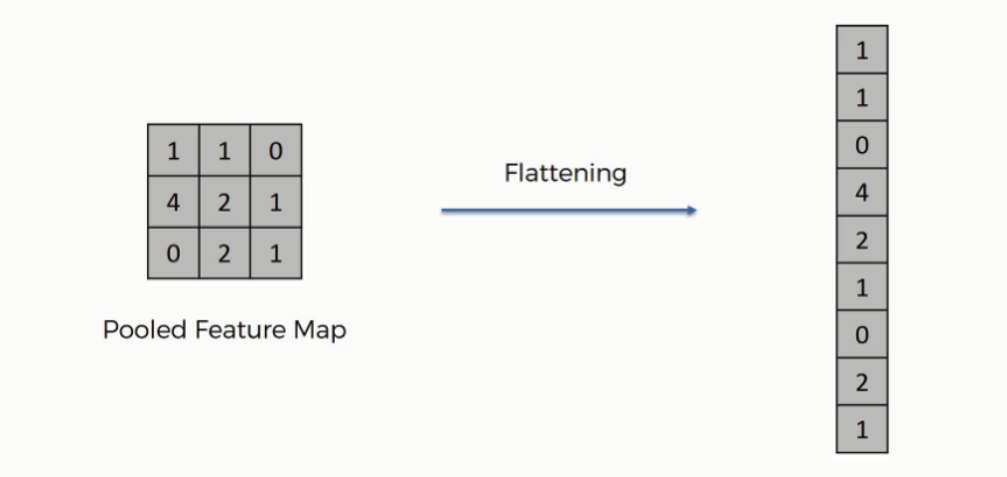


Fig. 13*: Flattening*

The vector of input data is then passed as input of an Artificial Neural Network and processed further. The penultimate layer, called Fully connected Layer, provides a vector of dimension K, where K is the number of classes that the network will be able to recognize. Finally, we have a **Fully Connected Neural Network**. This final layer uses a classification method to provide the output of the classification, i.e., the indication of the class to which the object represented in the input image belongs. The entire process is showed in the image below.

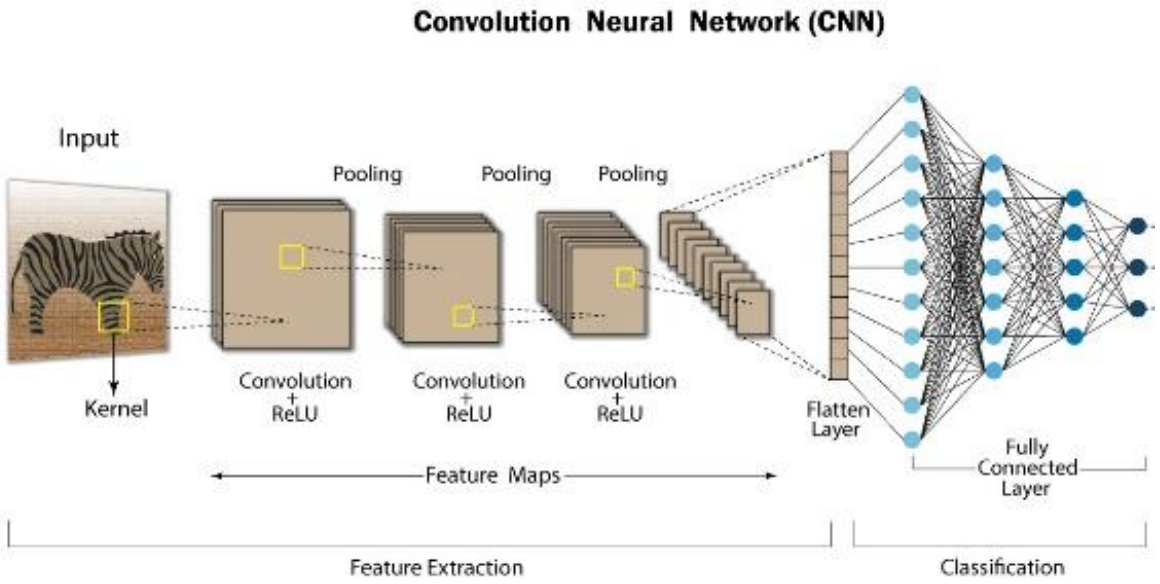


Fig. 14: *Example of CNN with multiple convolutional layers. The filters are applied to each training image with different resolutions, and the output of each convolved image is used as an input for the next level.*

Chapter 2

Applications of Artificial Intelligence in Medical Diagnosis

2.1 The benefits in the application of AI to Medicine

Among the different fields of application of AI, as discussed in Chapter 1.3, medicine is one of the most important and life changing so far. The implementation of Artificial Intelligence in healthcare simplifies the life of the medical staff and allows to speed up the diagnostic process through the number, images and words analysis made automatically by the machine. Other than speeding up diagnoses, it can improve the dialogue between doctor and patient, and it can support the treatment assigning an ad-hoc drug therapy. Additionally, the use of Artificial Intelligence in healthcare can accelerate scientific research, saving many human lives also thanks to a faster and better access to thousands of scientific publications.

According to research carried out by Frost & Sullivan's, the AI market in healthcare will reach $6 billion by 2022, with an annual growth rate of 68%, generating savings of more than $150 billion. Global Market Insights calculates that through 2025 we will see an annual growth rate of 41.7%. In the medical imaging sector alone, one of the most promising for the application of AI, the prediction is that the AI market will experience 30% growth per year in the period up to 2025, thanks to improved computing power, learning algorithms and the availability of ever larger data sets. Moreover, the pandemic of COVID-19 has contributed to bring AI and medicine closer, leaving many questions unanswered, starting with the delicate issue of privacy and management of sensitive patient data.

Particularly important is the application of AI in **predictive diagnostics**: through the use and interpretation of data, early signs of certain diseases can be picked up to help doctors make more accurate diagnoses, with the goal of reducing errors and developing methods for individualized medical treatment. Among the fields that are attracting many investors there are **rehabilitation** – with machines able to learn from the physical therapist's exercises and then replicate them on the patient – and **medical imaging**: for instance, Artificial Intelligence can support the decision-making process of radiologists, not only in the detection of the disease, but also in the image preparation and planning tasks. An increasingly crucial contribution is the one that AI offers to the so-called **precision medicine**, which is increasingly emerging as the medicine of the future. In fact, Machine Learning makes it possible to develop personalized predictive models, with the possibility to customize treatments instead of using a one-size-fits-all approach. In the next few pages, a more detailed description of the applications of AI today will be provided, trough use cases such as: **prostate cancer imaging, intelligent diagnosis of COVID-19** and **detection of breast cancer**.

2.1.1 Case 1: InnerEye by Microsoft for Medical Imaging AI

When we talk about images in medicine, we are talking about reports such as X-rays, CT scans, and MRIs that help the doctor confirm his treatment hypotheses. Beginning with image analysis, Microsoft developed InnerEye, **a system that can determine whether the patient has a tumor**. InnerEye is able to process visual information from an infinite library of scanned photos, making this an important case of new technology application in the medical field. The project is being worked on by the Health Intelligence team at Microsoft Research in Cambridge (UK), which will provide open software and a deep learning library for training and deploying models on medical images.

Microsoft's InnerEye Artificial Intelligence is built on an algorithm that learns from the plates viewed by doctors, stores all the information, and formulates a diagnosis in a very short time, cutting the expenses of disease detection while also saving the patient and the physicians' valuable time. Addenbrooke's Hospital in Cambridge will become the first hospital in the world to use the InnerEye Artificial Intelligence technology, developing a model that leverages the hospital's data to automatically highlight tumors on patient scans. Thus, an important application of InnerEye is **medical imaging**, to assist clinicians for image preparation and planning tasks in radiotherapy cancer treatment, monitor tumor progression, radiology or surgery planning and magnetic resonance. According to Microsoft research published in JAMA Network Open, AI can help clinicians execute radiation planning 13 times faster.

Image-Guided Radiotherapy for Head, Neck and Prostate Cancers

Without the assistance of an AI system, planning radiation can be a lengthy process: it begins with 3D imaging scans, which are made up of stacks of 2D images grouped in dozens of layers, each of which must be inspected by an expert for targeting the troublesome body region. The expert must manually draw a contour line around the tumors in each image: this method is known as "contouring", and it can take several hours or days for a single patient in difficult circumstances, consuming a significant amount of time and resources during cancer treatment while also increasing the hospital's overall costs. The implementation of InnerEye for picture segmentation is a feasible solution for this lengthy procedure, making it up to 13 times faster than doing it manually, and the accuracy is comparable to that of human specialists. The most recent applications include the use of Machine Learning models for radiation in malignancies of the head, neck, and prostate.

The utility of ML in radiotherapy was assessed by comparing the time required for manual image segmentation by clinicians with the time required when clinicians use ML models to assist them in contouring the images, while also accounting for the differences in interpretation between the experts and the model. The research is applied to image data sets of the prostate, head, and neck, and the code base for the deep learning training framework and segmentation models is publicly available on GitHub for reproducibility on different datasets.

Data Description

Pelvic CT Dataset: The model was applied to CT scans from seven institutions for male patients undergoing radiotherapy treatment for prostate cancer from 2012 to 2019. Such scans underwent quality and integrity checks regularly, and all the patients received radiation therapy to their intact prostate. Machine learning models were trained for prostate cancer with 519 images.

Head and Neck CT Dataset: all patients received radiation therapy for tumors of the upper aerodigestive tract for curative intent, and they were composed 24% by females and 76% by males. Machine learning models were trained for head and neck cancer with 242 images.

The pelvic and head/neck CT imaging datasets were collected at eight different clinics around the world, including Europe, Australia, New Zealand, North America and South America. This was important to enhance the **generalization of the segmentation models,** in order to make sure they work on unseen data and on different cancer types and body types.

Contour ground truth

Seven anatomical structures of the pelvic were contoured by clinicians to provide a ground truth. Such clinical contours presented inconsistencies among the specialists because of the different clinical protocols and accuracy the clinicians were aiming for. Therefore, to provide integrity, all contours were checked by expert radiation oncologists with more than ten years of experience. Differently, for Head and Neck CT Datasets, nine anatomical structures were contoured, showing much less intra-observer variation and more consistency: nonetheless, the contours were checked by experts anyways so to provide precise results.

Methodology

The CT volumes were normalized and resampled to a fixed image resolution for the training, while at test time full volumes were used as input samples. During training, the resampled CT volumes were taken randomly, and mini batches were created for each volume.

The segmentation model is based on a specific Convolutional Neural Network called "**3D-UNet architecture"**, useful for classifying labels used in the medical sector. However, in medical imaging, the desired output is more than a classification: such CNN also contains the localization that is used to predict the class label of each pixel, by providing an **output segmentation map** for analyzing the local region [Figure 15].

The Convolutional Neural Network takes on 39 million training parameters, composed by a sequence of convolution blocks to which is applied a ReLU activation function, encoding a given input 3D CT scan in multiple image scales to extract the necessary information for the segmentation end task.

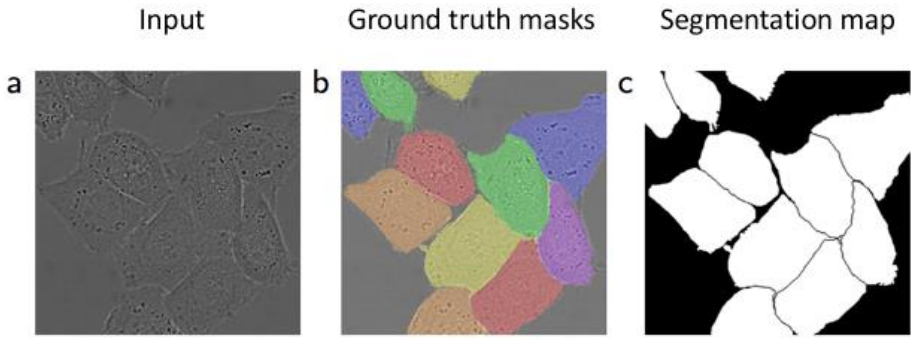


Fig. 15: *Segmentation map*

In the case of InnerEye, individual components of the segmentation model are the encoder, decoder and aggregation blocks, that finally lead to the output segmentation, as showed in Figure 16. The design used provides slight modifications to the traditional 3D U-Net architecture, so to better utilize the computation hardware and increase the number of samples used in model parameter updates in stochastic gradient descent.

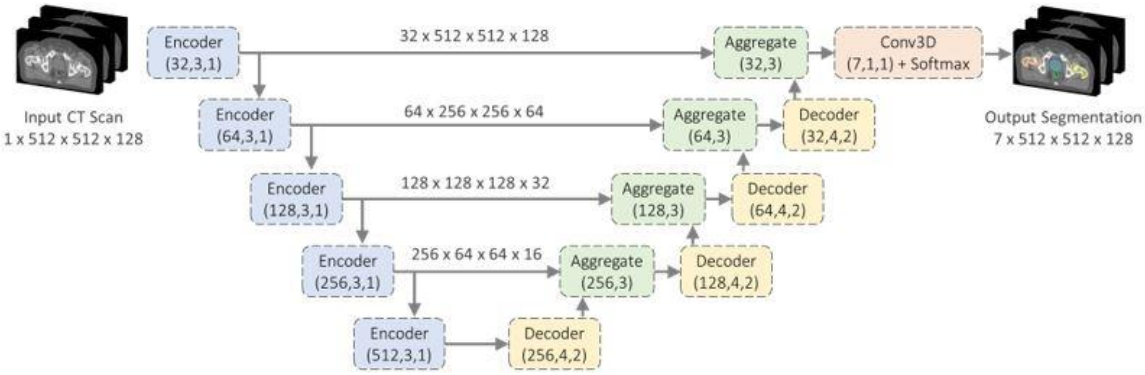


Fig. 16: *Image Segmentation model*

To decide the clinical applicability and the performance of the model, the previously trained pelvic and head/neck CT models have been tested on their corresponding external data set from three clinical sites that had different imaging protocols and CT scanners on patient preparation, showing that the algorithm performs consistently on main sites and external sites, increasing the confidence on the generalizability.

In Figure 17, we see the qualitative results on three different subjects: the first column shows how the algorithm performs on the main data set, while the second and third column show the algorithm

performance on an external data set. The darker colors represent the algorithm predictions, while the lighter colors represent the ground-truth annotations.

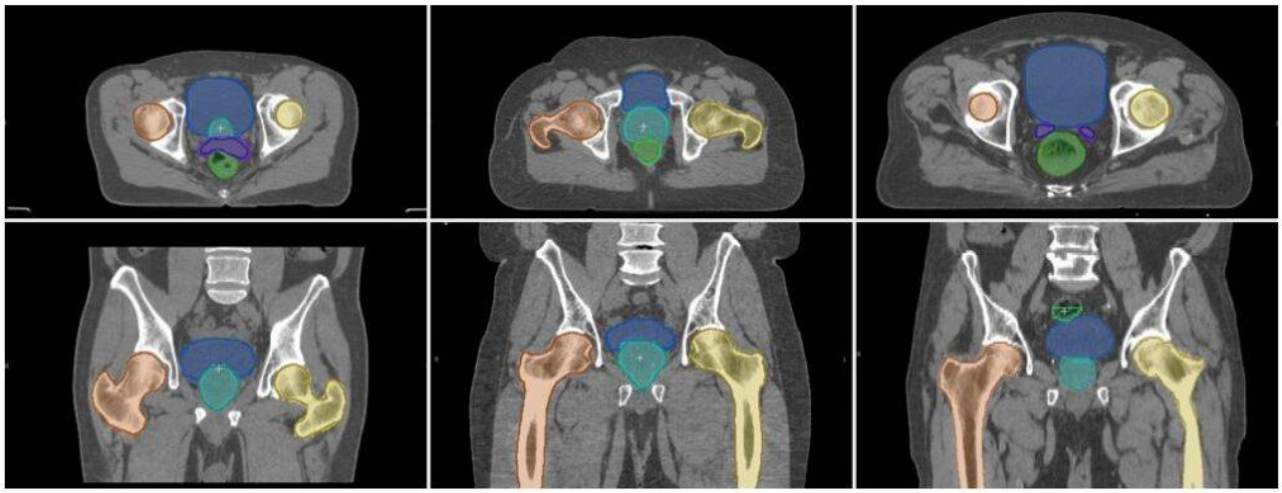


Fig. 17: *Comparison between algorithm and ground-truth on scan*

Evaluation Metrics

To evaluate the performance, one of the techniques used was the similarity metric "Dice coefficient", which compared the overlap between segmented structures in pairs of images. Perfectly overlapping structures, and hence perfect agreement, results in a score of 100%, while a score of 0% corresponds to complete lack of overlap. Also, to assess the agreement between generated contours and clinicians' ones, statistical measurements such as Cohen's Kappa and Feliss' Kappa were used, measuring an agreement score on the pixels for each structure. The performance differences between the main and external data sets were tested with the Mann-Whitney test, useful for understanding if the model performance drops because of the generalization issues.

Results

No inconsistencies were found that, if not corrected, could lead to significant errors in a treatment. Indeed, the prostate segmentation results showed that the algorithm's organ delineations of prostate scans were consistent with the contours delineated by the clinicians, and their error stood within the acceptable error bound. Also head and neck segmentation results were consistent, and the model performed well on validation and external data sets.

Looking at the clinical utility aspect and measuring the clinical value of the mode in terms of time reduction, for head and neck the time can be reduced from 73 minutes to 5 minutes, which is more than 90% time savings. In this way, the clinicians can gain more time to spend with their patients

for person-to-person interactions. In the following image, the blue bar represents the time taken to perform the segmentation scan manually by an expert radiation oncologist on then head and neck CT scans with varying imaging quality. The bottom bar shows the model's runtime plus the time for the same experts to review and update the contours while ensuring the clinical accuracy. Similar gains can be observed also for the prostate scans.

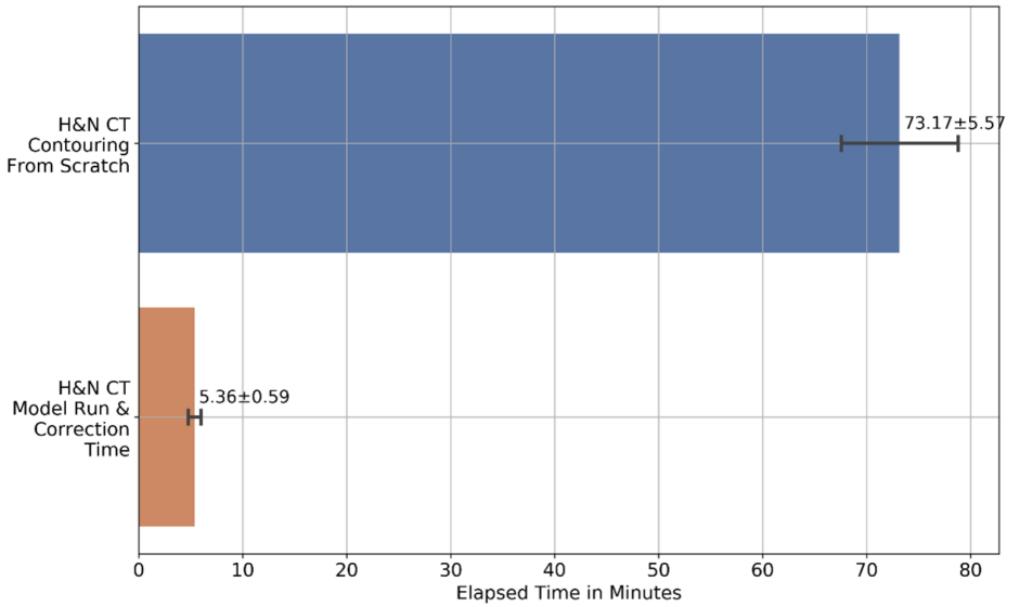


Fig. 18: *Annotation Time Savings*

The time reduced does not only involve the time taken to draw segmentation contours on the image, but also the time to correct inaccuracies of the automated system.

Reduced planning time is an important utility of these models, however, this should be complemented by clinically acceptable contours generated by the AI as well. To assess this, a separate experiment was carried out together with multiple clinical experts to measure the inter-expert variability on this contouring task on a subset of images. The manual annotations were collected by a set of radiation oncologist experts, from each clinician per structure, creating a gold standard of the contours extracted from an aggregation of multiple segmentation maps. Afterwards, the variability across these experts was observed to understand the bounds of clinical acceptability, and to perform a statistical agreement analysis between the contours generated by the clinicians and the algorithm.

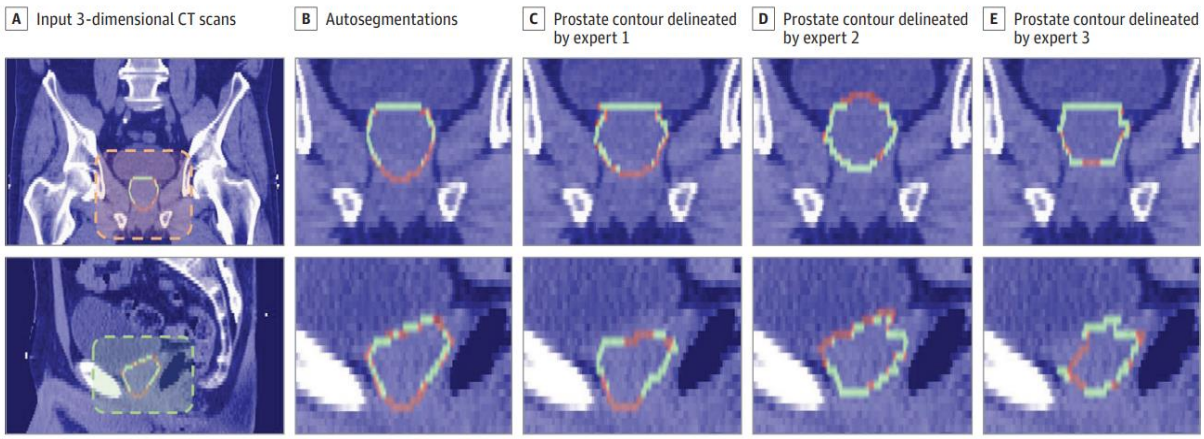


Fig. 19: *Inter-Expert Variability in Prostate Contour Annotations*

In the image above, the prostate contours collected from the model and the experts are shown, obtained from two different pelvis scans. The color map shows the surface distance areas with respect to the gold standard contours: the red color represents the surface points with distance error larger than 2.8 mm. It is clear to see that the model contours are aligned with the ones originating from the experts: the model does not perform arbitrary segmentation and its error is consistent with the expert's errors.

Conclusions

InnerEye has shown to the clinical world that ML models can be integrated into traditional radiotherapy practices since they are consistent with the bounds of human expert variability. The algorithm's autosegmentation reduces contouring time while yielding clinically valid structural contours on heterogeneous datasets for both prostate and head and neck radiotherapy planning. Also, the model robustness is tested through images from clinical sites coming from all over the world, with their respective protocols and hardware. The use of InnerEye saves 90% of clinicians' time, and the platform has an underlying cloud technology that can be used for the integration into existing clinical workflows.

Whereas the ML models perform well, it is crucial that the clinicians keep assessing their accuracy and clinical significance: this is a necessary component for the correct functioning of the ML-augmented radiotherapy workflow. At the same time, clinicians can inspect and edit contours in minutes rather than hours. The source code used in this study is publicly available: this creates an opportunity for oncology centers to use the technology to train and deploy new models with their own data sets.

2.1.2 Case 2: Artificial Intelligence and early diagnosis of COVID-19

The spread of the COVID-19 pandemic triggered the scientific community an immediate response to face the virus, seeking adequate treatments and vaccines for the prevention of the disease. Since the new virus was already present in China at the beginning of the pandemic, the diagnostic techniques of the illness were already published by the Chinese and hence they appeared easy to reproduce, with the approach of nasopharyngeal swabs, serological and antigenic tests, chest radiographs and other methods that were not novel to the scientific community. The known sequence of the virus led to the fast development of vaccines with the technologies such as mRNA, but what seems less developed at present are the techniques for early diagnosis of the disease, which would be of great utility especially in cases of interstitial pneumonia which is the main cause of ICU admissions and deaths.

The aim of this research, published by the Department of Electrical and Information Engineering of the Polytechnic University of Bari, Italy, is to show how the technologies employed in electronics for the acquisition and processing of data could allow a precise and early diagnosis of the onset of COVID-19 interstitial pneumonia. For this purpose, at first screening the use of smartphones might be sufficient, avoiding expensive medical equipment and diagnostic tests that are prohibitive and burdensome for waiting times. Therefore, we will first see how it is possible from a medical point of view to diagnose the onset of interstitial pneumonia early; subsequently we will make an overview of the algorithmic methods useful for this purpose, and finally we will see how a smartphone can be a useful tool for early diagnosis.

Introduction

Lungs are two organs responsible for the supply of oxygen to the body and for the elimination of carbon dioxide from the blood, that is the gaseous exchange between air and blood (a process known as hematosis). Located in the thoracic cavity, lungs are enveloped by a serous membrane, the pleura, which is essential for the performance of their functions. The lungs are separated by a space between the spinal column and the sternum, the mediastinum, which includes the heart, esophagus, trachea, bronchi, thymus and large vessels. Their main task is to receive the blood laden with carbon dioxide and waste products from the peripheral circulation and to "clean" it, enriching it with oxygen and then sending it to the heart, from where it is spread to organs and tissues. Physiological respiration is an automatic process, unconsciously controlled by the respiratory center

at the base of the brain. Sensory organs located in the brain, aorta and carotid arteries monitor and detect blood levels of oxygen and carbon dioxide within the body.

In healthy individuals, an increase in carbon dioxide concentration is the most important stimulus for deeper and faster breathing. Conversely, when the carbon dioxide concentration in the blood is low, the brain reduces the frequency and depth of breathing. During breathing it is possible to acquire a signal [Figure 20] whose morphology is similar to a sinusoid of period between 3.3 and 5 s having a frequency between 0.2 and 0.3 Hz in a healthy subject, whose breathing can vary between 12 and 18 acts/min.

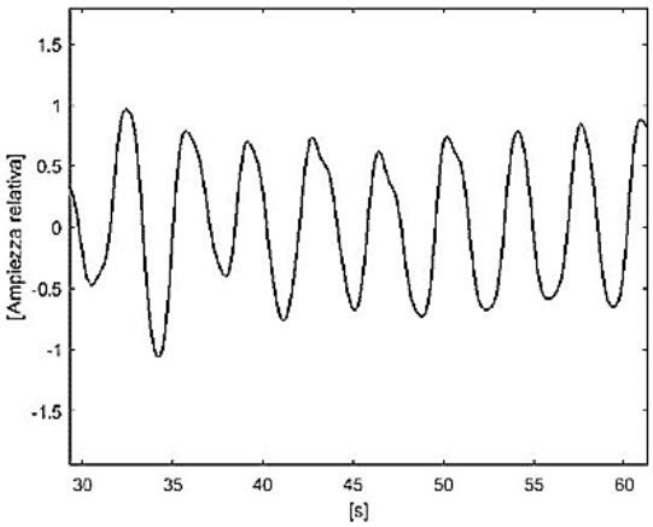


Fig. 20: *Respiratory signal*

In this signal, ascending curves represent inhalations and descending ones represent exhalations, whereas the depth of the breath depends on the patients and the activity they are performing. In fact, depending on the type of breathing and the state of the individual, the signal can have different amplitude and the peaks can be more or less close in time. Therefore, from the respiratory frequency and the morphology of its signal it is possible to trace back to possible pathologies, such as COVID-19. The presence of any respiratory disease manifests itself as an alteration of the sounds heard by medical experts in amplitude, duration and frequency.

Furthermore, there is a specific signal generated by the respiratory system that can provide vital diagnostic information, and that signal is the cough sound. Specific sound qualities can aid in the early detection of life-threatening diseases like pneumonia. Artificial Intelligence is currently widely and consistently used to recognize signals, objects, and faces, as well as to define signal properties. The basis of the COVID-19 pneumonia early detection system is to convert respiratory

sounds (including coughing) into images, which are then classified as normal or pathological using AI. This classification is the exact diagnosis required, and it can be accomplished with a simple smartphone.

Methodology for the Processing of Audio Signals for Classification with AI

To produce an early diagnosis of COVID-19 from the cough sound, Deep Learning algorithms are more indicated because of the automaticity with which features are identified and extracted, but also simple Machine Learning could be used if there are not enough data available to train the Artificial Neural Network. From the overview of the AI methods described in Chapter 1, we can conclude that it is possible to carry out recognition of images through Deep Learning algorithms, by extracting features or characteristics in order to execute the classification process. In our case, in order to create and train the Deep Learning model, the **Transfer Learning (TL)** approach is implemented on Convolutional Neural Networks.

The process consists in the modification and re-training of an already existing model that has been previously trained (pre-trained net) to recognize objects categories (classes) different from those of interest. The point of start is hence an existing neural network (such as AlexNet, GoogleNet, SqueezeNet etc.), which is modified and re-trained to recognize new classes of objects. In contrast to beginning from scratch, this approach exhibits the benefit of not needing to create an ANN from the start and requiring significantly less data for training; thus, computation time is significantly reduced. It is necessary to transform audio into images before using this procedure on audio data, hence, it is vital to describe how this transformation might be accomplished effectively.

An audio signal is a pressure wave propagating in a transmission medium such as the air. A method of processing of audio signals, very useful when applied in conjunction with AI models, is the **Fourier Transform (FFT)**. The FFT is a mathematical tool through which it is possible to **convert a time-varying signal into a time-frequency representation**. This representation describes the spectral content (i.e. frequency) of the signal itself, and how it evolves over time, generating graphs known as "**spectrograms**" and "**scalograms**". From these graphs, or images, are extracted the features for the recognition by the AI.

The spectrogram is a graphical representation of the intensity of a sound as a function of time and frequency. It is therefore a graph that shows the frequencies that make up the sound wave as time passes. The spectrogram contains information about the amplitude of the wave (and therefore the intensity of the sound), expressed through a color code.

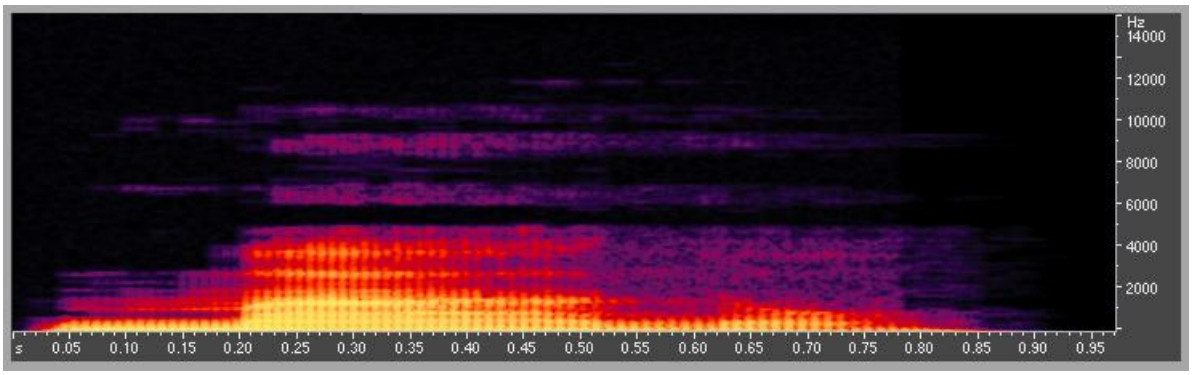


Fig. 16: *Color spectrogram of the word "hand" pronounced by an Italian native speaker. Lighter colors (yellow) indicate a higher sound intensity than darker colors (purple)*

In the image above, the ordinate axis (vertical, y-axis) shows the frequencies, while on the abscissa axis (horizontal, x-axis) the time. The presence of multiple horizontal lines, i.e., multiple frequencies, indicates that it is not a pure sound (also known as “mono-frequency”).

Speaking of scalograms, they are the representation of a wavelet using an oscillating waveform of finite length or of fast decay, scaled and translated to fit the input signal and localized in both time and frequency. The scalogram is obtained by sampling the signal with a window length of constant duration which is shifted in time and frequency, unlike the spectrogram in which it is fixed. To calculate a scalogram it is first necessary to sample the signal in overlapping segments and, for each of them, calculate the constant wavelet coefficients.

Compared to FFT, the Wavelet is computationally less expensive (O(N) instead of O(NlogN)), and the scalogram may be more useful than the spectrogram, for example, for analyzing slowly varying signals punctuated by sudden transients. An example of a scalogram is shown in Figure 17.

In this research, six techniques for transforming audio signals into images are compared with each other in order to determine which one is the most suitable for the classification of the signal produced by the COVID-19 cough, and to distinguish it from coughs caused by other pathologies. These techniques are in some respects similar to each other, in that they indicate multiple ways of representing time-frequency of an audio signal; however, they present differences that recommend for each a specific field of application.

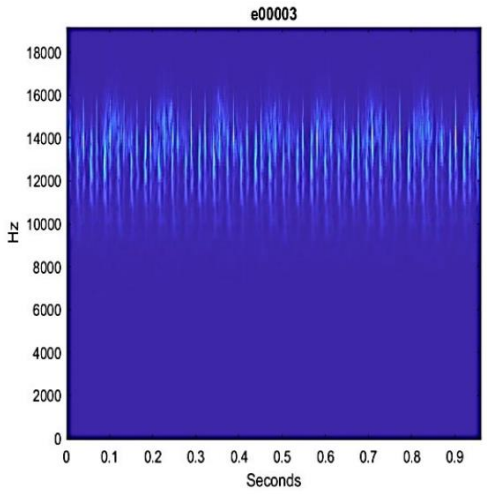


Fig. 17: *Example of scalogram for cardiac audio signal*

The methods for transformation of audio signals into images implemented and compared with each other are:

- **Mel Spectrogram:** the frequencies are converted into the "Mel" scale; this type of technique is based on the comparison of sound pitches.
- **Gammatone Spectrogram:** very similar to the Mel spectrogram, with the only difference that it is calculated on a different scale, called ERB (Equivalent Rectangular Bandwidth).
- **Continuous Wavelet Transform (CWT):** provides an overcomplete representation of a signal by letting the translation and scale parameter of the wavelets vary continuously.
- **Mel-Frequency Cepstral Coefficients (MFCC) image:** from the Mel spectrogram of an audio signal we can extract the characteristic elements, or features, known as Mel-Frequency Cepstral Coefficients (MFCC). They allow a representation of the real cepstrum of a signal, which is the result of the Fourier transform applied to the spectrum of a signal, often called "spectrum of the spectrum".
- **Image of GTCC coefficients (GTCC):** the Gammatone spectrogram and the relative cepstral coefficients at the Gammatone frequency (GTCC), are very interesting because they are less vulnerable than other types of spectrograms with noise components superimposed on the signal.
- **Image of CWT coefficients:** characteristic features of the Continuous Wavelet Transform representation.

The Mel spectrogram and its MFCC coefficients are among the most widely used techniques because they are tailored to human auditory sensitivity. However, they have limitations related to

the low efficiency of Mel filters in eliminating additive noise, especially the one found in speech audio signals, and therefore in speech recognition applications. The problem of noise in speech recognition applications can be solved by resorting to the Gammatone spectrogram and its GTCC coefficients, which better filter the frequencies in which the additive noise is mainly located. The GTCC and its spectrogram Gammatone are thus more suitable for voice recognition and identification, whereas the MFCC and its spectrogram Mel are more suitable for classifying general audio signals that are not very noisy, such as heartbeats and biological signals in general, where the time-varying behavior is well defined and sampling is easier, as well as considering that additive noise is much more mitigated because they are assumed to be acquired using a dedicated hardware with low noise.

The wavelet approach is useful to process information with higher resolution compared to other techniques, because the period of analysis is not fixed, and hence allows one to better capture audio signals with long time intervals at low frequencies, or, on the contrary, very short temporal intervals at high frequencies. This technique seems therefore particularly indicated for the classification of respiratory signals where we detect, especially if pathological, sounds that overlap the vesicular murmur.[1]

Comparisons and Results

The images obtained from the 6 techniques were classified using multiple CNNs in order to determine not only the most effective method of transforming the audio signal into images for accurate classification, but also for understanding the type of pretrained network able to provide better accuracy using the TL technique.

The first database used is of 3240 audio files of heart sounds (so-called cardiac tones), acquired through the technique of phonocardiography (PCG), converted then into images and subsequently given as input to the neural network both for training and test. Of these audio files, 2574 are related to healthy heart tones and 666 to pathological heart tones. Subsequently, using a database of both heart tones (817 healthy and 183 pathological) and respiratory sounds (35 healthy and 885 pathological), a comparison was made of the performance between various pre-trained networks customized with TL technology, transforming audio files into spectrograms and scalograms.

The first pretrained network used and retrained on Matlab according to the TL technique is GoogleNet, where a few layers have been modified before proceeding with the training. The first

[1] characteristic noise, due to the penetration of air into the pulmonary alveoli, which is perceived in healthy subjects on auscultation of the chest.

layer has been changed because the default image dimension of the *imageInputLayer* had dimension 224x224x3, instead the images to analyze have dimension size 227x227x3; after that, the output of the *fullyConnectedLayer* has been modified by setting it to 2, as the output classes are two, normal and abnormal (pathological). For the same reason, also the last layer *ClassificationLayer* has been modified. Finally, the network has been trained setting a value of Learning Rate equal to 0.0001 and number of epochs equal to 6. This procedure has been applied for all the 6 methods of the transformation of the audio signal into images previously described.

From the results obtained, a first comparison can be made between the techniques related to time-frequency representations: Mel spectrogram, Gammatone spectrogram and scalogram. Below, are shown the images related to the same audio file resulting from the acquisition of heart tones transformed according to each of these three techniques.

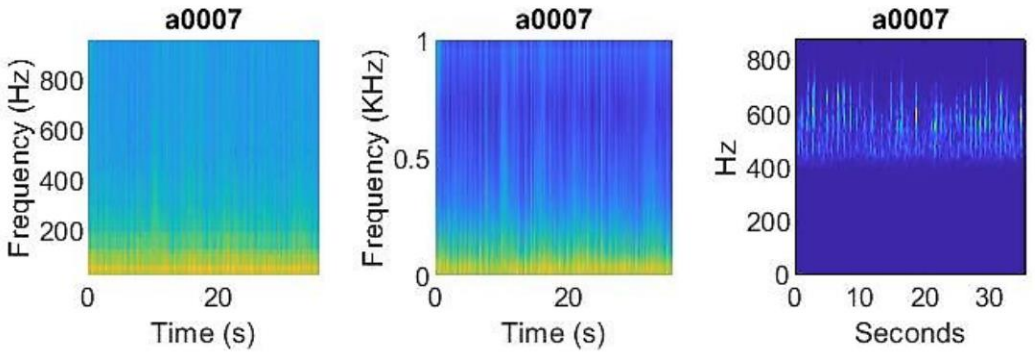


Fig. 18: *Mel Spectrogram, Gammatone and scalogram of healthy cardiac tone*

The Mel spectrogram (on the left) better portrays the tone of the heartbeat, even at low frequencies, which are the most crucial for successful categorization because they have more information content.

A further comparison can be done by extracting the features from the Mel Spectrogram, Gammatone Spectrogram and Scalogram, generating MFCC, GTCC and wavelets coefficients, subsequently expressing them as images, as shown below.

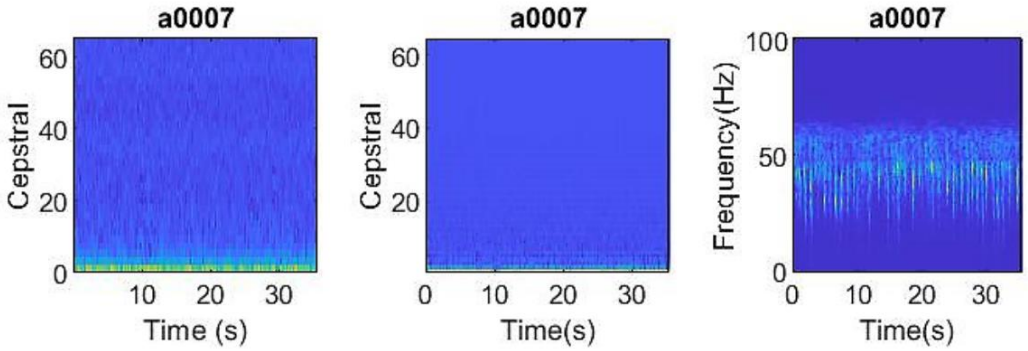


Fig. 19: *MFCC, GTCC, CWT coefficients of healthy cardiac tone*

It is possible to notice that the image obtained from the MFCCs (left) shows more evident tonal variations, thus being more suitable for DL classification. Also, the image obtained from the wavelet coefficients (on the right) shows a good color tone variation, while the less marked variations are those present in the image obtained from the GTCC (middle).

The table below compares the performance, in terms of accuracy and loss for each epoch, of the CNN based on the use of GoogleNet and trained with the TL method, relative to each of the techniques for transforming sounds into images.

TECHNIQUE	ACCURACY	LOSS
Mel Spectrogram	**93%**	**1.1**
Gammatone Spectrogram	86.63%	1.5
Scalogram	89%	3.5
MFCC	**93%**	**1.2**
GTCC	91%	1.3
Wavelets	90%	1.7

Table 1: *comparison of techniques implemented with GoogleNet network re-trained with TL*

As we can observe from the table and as already anticipated through qualitative analysis, in the classification of the cardiac tones the technique relative to the Mel spectrogram presents a higher value of accuracy than the other two techniques and a lower value of loss. Therefore, among the three techniques relative to time-frequency representations, this is the most accurate and reliable. among the techniques for feature extraction, the cepstral coefficients at the Mel frequency (MFCC) appear to offer greater accuracy than the other two techniques (GTCC and coefficients of the CWT).

Subsequently, using a database of both heart tones (817 healthy and 183 pathological) and respiratory sounds (35 normal and 885 pathological), a comparison of performance between various

pre-trained networks was carried out, transforming the audio files into spectrograms and scalograms. The networks compared in Matlab environment were: GoogleNet, SqueezeNet, AlexNet and ResNet50. The performance comparison is shown in Table 2.

CNN (Convolutional Neural Network)	Validation accuracy Mel spectrograms for heart rate	Validation accuracy scalograms for heart rate	Validation accuracy Mel spectrograms for respiratory sounds	Validation accuracy scalograms for respiratory sounds
GoogLeNet	**99.5%**	97%	95.65%	**100%**
SqueezeNet	98.5%	98.59%	96.2%	100%
ResNet50	95%	99%	96.74%	100%
AlexNet	96.33%	97.67%	96.36%	100%

***Table 2**: Performance comparison between different pretrained and retrained networks using the TL method to classify two types of biological sounds: heart rates and respiratory sounds*

The highest accuracy (99.5%) regarding the heart tone dataset, was found in the GoogLeNet network. In the case of the respiratory sounds dataset, the highest accuracy (100%) was found in all 4 networks.

It is very clear from the results obtained that the **Mel spectrogram and the coefficients MFCC** are the sound processing methods that best suit AI algorithms to recognize, with a high degree of confidence, biological sounds from the respiratory system by classifying them as healthy or pathological, and that the GoogleNet network is the pre-trained CNN, among those considered, that offers the highest degree of confidence in the classification. In order for this procedure to be applied to the early diagnosis of pneumonia by COVID-19 it is necessary to collect a set of audio files that is the largest possible, so to apply the TL method for fine training of one of the CNNs — presumably the GoogleNet — for cough recognition generated by COVID-19. Preliminary studies on a limited number of samples demonstrate an accuracy of no less than 80 percent which has potential margins for increase to at least 90% and beyond.

Smartphone app development

The CNN trained via TL can be implemented in AI models executable on both PCs as well as on smartphones, thus allowing the best exploitation for diagnostic purposes of smartphones' computational and graphical potential. To develop apps that can be executed in both the Android and iOS environments, one must obviously be an expert programmer; however, at least for a

development aimed at verifying the idea and the debugging of the algorithm, there are convenient code translation tools that allow high-level programming, often visual thus quite simple, that do not require particular in-depth knowledge of programming languages. This is the case with the Matlab/Simulink environment. In particular, Simulink offers a very interesting and useful smartphone app development tool, starting with the design of a functional block model (thus not directly written in a programming language) which is then automatically translated and converted into an executable app on smartphone, as described below.

A. Importing the Trained Network

There are two ways in the Simulink environment to implement AI algorithms: one is by using the *Image Classifier* functional block, in which one imports a CNN trained with TL from the neural network project and then exports it as a model; the other way is the use of a customized Matlab functional block in which the code is inserted to execute a CNN previously trained and exported as a *Compact Model*. As input to the functional block, the image to be classified is provided (spectrogram, scalogram, image from features). The output is the classification (prediction) of the input data provided i.e., healthy/pathological sound signal in our case. Obviously, in order to classify sounds, it is first necessary to acquire them.

B. Audio Acquisition and transformation in image

The audio acquisition of the biological signals can be done through the microphone of a smartphone or, if this is not of adequate quality, an external microphone connected to the smartphone can be utilized. Each audio file to be classified is converted into images via an additional functional block specially developed in Matlab and inserted into the Simulink model: based on the previous considerations, the use of Mel's spectrogram and the extraction of the MFCC coefficients for heart tones, and the use of the scalogram with the CWT coefficients for respiratory sounds are the most suitable.

C. Classification with CNN

Finally the CNN is trained with the TL and exported in an *Image Classifier* block i.e., in a *custom Matlab Function* type block at completion of the model, which is then translated into app installed and executed on smartphones. This provides an app that can automatically and continuously monitor the state of health and facilitate early diagnosis of COVID-19 from the first strokes of the cough.

Conclusions

The study aims to demonstrate how AI tools can be applied for one of the most important treatment goals of COVID-19: the early diagnosis of interstitial pneumonia. Such diagnosis would allow a reduction in hospitalizations, an increase in the probability of survival by being able to undertake timely appropriate therapies, and a reduction in the risk of infection, where even pauci-symptomatic subjects would be detected early. The possibility in terms of engineering tools available for the purpose has been demonstrated, being moreover tools that could be easily implemented on smartphones via dedicated apps that would have a very affordable cost.

For this study to be applied and succeed, the next step is to have a large database of recordings of coughs from patients with different diseases, including COVID-19 of course, because it is essential for training the neural network and to achieve levels of classification confidence that are as high as possible, tending toward 100%. For this it will be necessary to organize coordinated projects between teams of physicians and engineers.

2.1.3 Case 3: Detection of Breast Cancer with Mammography through AI Support System

According to data from the report "The Numbers of Cancer in Italy 2021", in Italy there were 55,000 new diagnoses of female breast cancer in 2020 and an estimated 12,500 deaths in 2021. Breast cancer is the most common malignancy among women, and it is the leading cause of cancer death in the female population. Such threat can be limited by using Artificial Intelligence systems in support of physicians during the diagnosis, to improve cancer detection at mammography and boost the diagnostic accuracy and the recall rate. AI systems can also reduce the radiologists' workload and patients' unnecessary examinations, radiation exposure and stress. Mammography screening for breast cancer is thought to be useful in lowering breast cancer-related mortality. However, given the enormous number of women checked, a high workload threatens efficiency, especially given the paucity of screening radiologists. Furthermore, it is critical to reduce interpretation mistakes, which account for at least 25% of missed detected malignancies.

Computer-assisted detection (CAD) technologies were developed to help radiologists enhance their detection ability. Few studies have examined the actual benefit of single reading plus CAD over single reading alone (i.e., the actual impact of radiologists in screening performance). In general, the value of CAD in screening remains uncertain. Most research suggest that the cost-effectiveness of screening has not improved, owing to the low specificity of most traditional CAD systems.

However, advances in Artificial Intelligence with deep Convolutional Neural Networks are lowering the performance gap between humans and computers in a variety of medical imaging applications, including breast cancer diagnosis. As a result, this new generation of **deep learning–based CAD systems** may ultimately enable breast cancer screening programs to enhance their performance. The purpose of the following pages is to evaluate radiologists' breast cancer diagnosis ability when reading mammographic images without assistance versus when using a commercially available AI system, in terms of overall diagnostic performance and efficiency.

Materials and Methods

The study used digital mammographic pictures from screening exams that have been anonymized and collected subsequently. Women from two institutions were included: one in the United States (collection center A) and one in Europe (collection center B), targeting a dataset of 240 digital mammographic examinations (100 showing cancer, 40 false-positives, 100 healthy). Women who came in for screening with no symptoms or concerns were the only ones that were included in the study. Women who had implants or had a history of breast cancer were excluded from the research.

AI Support System

For the purpose of the research, the radiologists relied on ***Transpara***, an AI computer system for assistance (version 1.3.0, ScreenPoint Medical). In mammography and breast tomosynthesis, this technology is designed for **automatic breast cancer detection**. By using convolutional neural networks, feature classifiers and image analysis algorithms, the AI is able to depict calcifications and soft-tissue lesions. The findings of soft tissue and calcification are then combined to form **suspicious region findings.** Each location is given a number between 1 and 100, reflecting the level of suspicion on the presence of cancer (100 signifying the highest suspicion). Finally, the scores of all detected regions are combined to create an examination-based score (*Transpara score*), which ranges from 1 to 10 (with 10 indicating the highest likelihood that cancer is present on the mammogram).

The AI system is trained, validated and tested using a database of over 9000 mammograms with cancer (one-third of which are shown as lesions with calcifications) and a similar number of mammograms without abnormalities. The AI system is validated using an internal data set that has not been utilized for algorithm training or validation. Furthermore, the algorithms utilized in this work were never trained, validated, or tested using mammograms before.

In practice, by clicking on a specific breast region, **the radiologists can engage interactive decision-making support**. If something in that region has been identified, the system displays its

level of suspicion (on a scale from 1 to 100); otherwise, nothing is displayed, save a little cross marking the clicked place. Furthermore, the system always displays a **proprietary examination score** (*Transpara score*) between 1 and 10 on a whole-examination basis.

Evaluation

To examine both reading conditions — unaided and with AI support — a fully comprehensive, multireader, multicase evaluation was conducted in two sessions (at least 4 weeks apart). The evaluation took place in two locations (evaluation centers A and B), and it was carried out by three general radiologists and 11 breast radiologists.

During each session, radiologists read half of the exams with AI assistance and the other half without. Any information regarding the patient, including past radiography and histology reports, was kept hidden from the radiologists. Also, each specialist was individually instructed in a session with 45 examinations not included in the final evaluation before the first session. The goal of the training was to familiarize radiologists with the evaluation workstation, evaluation criteria and the AI assistance system (e.g., how to use all of its features). Radiologists assigned a mandatory **Breast Imaging Reporting and Data System** (BI-RADS) **score** (range 1–5) and a **probability of malignancy** (POM) between 1 and 100 for each examination (with 100 signifying extremely suspected malignancy).

Statistical Analysis

The study's major goals were to examine the **Area Under the receiver operating Characteristic** (AUC) curve, **sensitivity and specificity**, and **reading duration** between reading without assistance and reading with AI assistance. Secondary analyses were also conducted to gain a more complete understanding on the impact of applying an AI system to assist the reading of mammograms. The POM score was used to calculate ROC curves and associated AUCs.

The P-value was also computed[2] for rejecting the null hypothesis that readings conducted unaided or with AI help had equivalent performance. $P < .05$ indicates a statistically significant difference between the two reading conditions.

[2] The P-value was computed using *The Obuchowski-Rockette and Dorfman-Berbaum-Metz* mixed-model analysis of variance. It is the probability of obtaining a specific set of observations if the null hypothesis were true. When the P-value is less or equal than 0.05, the null hypothesis is rejected. When the P-value is greater than 0.05, the null hypothesis is accepted.

The BI-RADS scores were used to calculate sensitivity and specificity for each reading condition (i.e., with or without AI support). For each modality, the reader-averaged sensitivity and specificity were calculated.

Results

With AI support, radiologists increased their detection performance **improving the average AUC from 0.87 to 0.89** (difference, 0.02; P =.002) [Table 3]. The changes in AUC per reader ranged from 0.0 to 0.05 and were larger with AI support for 12 of the 14 radiologists (the other two readers had no change in AUC).

Sensitivity was 3 percentage points greater on average with AI support (P =.046), **specificity** was likewise on the rise (2 percentage points higher with AI support; P =.06), as shown in Table 4.

Figure 20 shows the ROC curves that demonstrate examples of tests in which the total number of correct recall assessments across readers varied between reading circumstances. In total, 32 examinations had a disagreement (i.e., at least three radiologists changed their assessments between reading conditions): In 72 percent of examinations (23 of 32), **the number of readers making the correct interpretation increased when AI support was used**, whereas the opposite occurred in the other 28 percent (nine of 32).

The reading time per case was similar in the unassisted sessions (146 seconds) and AI-assisted sessions (149 seconds). Nine of the 14 radiologists saw their reading time increase (range: 0.5%–10%) whereas five saw it reduce (range: 0.3%–22%). The computer ***Transpara score*** differed when reading alone and with AI support (P < .001). When **adopting the AI system**, radiologists **reduced their average reading time** per case **by 11%** for low-suspicion examinations (scoring 1–5). In contrast**,** using AI help increased reading time per case by 2% for high-suspicion tests (score, 6–10).

Assuming that each *Transpara score* category contains the same number of examinations in a screening population (and that examinations with a score of 1–5 account for half of the total and those with a score of 6–10 account for the other half), averaging the above-mentioned results should result in a **4.5% reduction in reading time per case when the AI system is used in screening**.

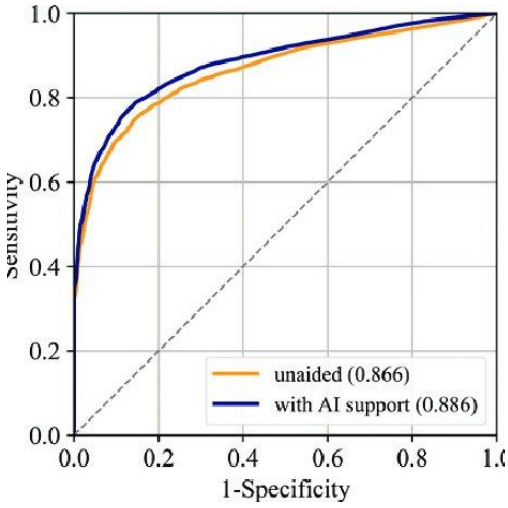


Fig. 20*: Average receiver operating characteristic (ROC) curves for two reading conditions: unaided and with AI support. The average is calculated across the 14 radiologists that took part in the study. Areas under the ROC curve are indicated by numbers in pare*

Table 3*: AUC for Each Radiologist and Reader-average AUCs for reading unaided vs aided.*

Variable	**Unaided**	**With AI Support**	**Difference**	**P-value**
Radiologist No.				
1	0.87	0.90	0.04	
2	0.82	0.84	0.02	
3	0.91	0.92	0.01	
4	0.85	0.85	0.01	
5	0.79	0.85	0.05	
6	0.84	0.86	0.02	
7	0.93	0.95	0.01	
8	0.87	0.90	0.04	
9	0.87	0.87	0.0	
10	0.90	0.92	0.02	
11	0.86	0.90	0.04	
12	0.86	0.86	0.0	
13	0.87	0.90	0.03	
14	0.87	0.88	0.01	
Average	0.87 (0.83, 0.90)	0.89 (0.85, 0.92)	0.02 (0.01, 0.03)	0.002

***Table 4**: Mean Sensitivity and Specificity across Radiologists*

Variable	Unaided	With AI Support	Difference (Percentage Points)	P-value
Sensitivity (%)	83	86	3	.046
Specificity (%)	77	79	2	.06

Conclusions

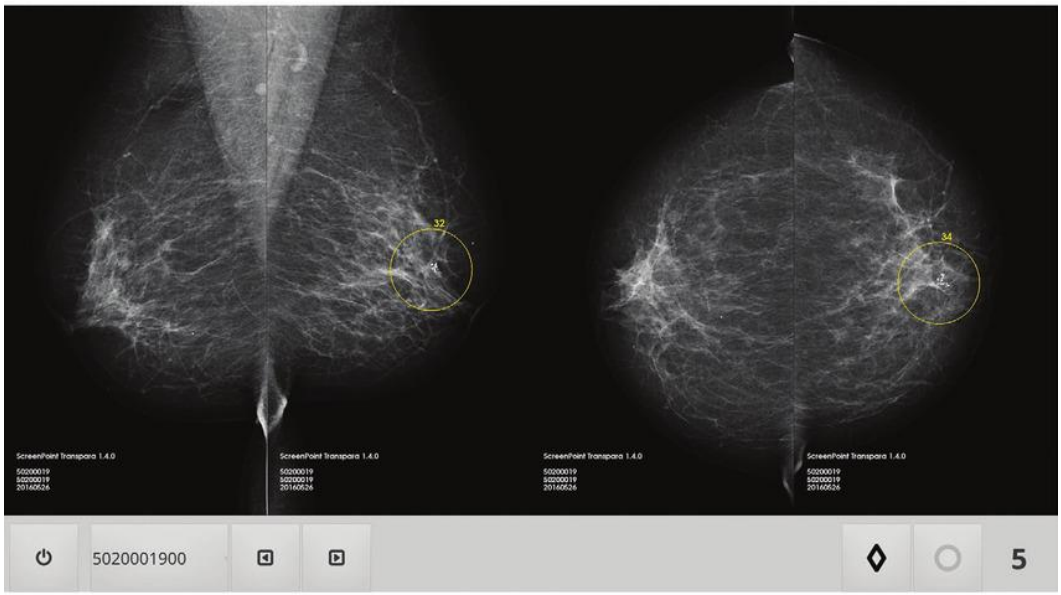
In the research, it has been discovered that **radiologists with AI support had better diagnostic performance** (as evaluated by the AUC) than those who read unaided. Under both conditions, the **average reading times per case were similar**. A cancer enriched data set of digital mammographic examinations with a representative sample of anomalies that may be seen in asymptomatic women receiving mammographic screening showed this improvement in diagnosis capability. A rise in the middle section of the ROC curve was responsible for the AI system's improved diagnostic performance. This shows that the AI aids in the evaluation of ambiguous instances, indicating the tool's clinical utility.

The radiologists with the least experience in mammography saw the **greatest improvement in AUC** with AI help. On the basis of experience, there was no difference in unaided performance. According to *Hupse et al*, some experienced radiologists may query the decision assistance more frequently: this might imply that they are less willing, or slower, to adopt new technologies or techniques to improve their performance.

Given the high workload of screening programs, the performance benefit of applying AI help is further boosted from a cost-effectiveness standpoint by the fact that **radiologists do not extend their reading time while using this system**. In fact, in a real-world screening scenario, the average reading time per case would drop by about 4.5 percent.

This suggests that the system's examination-based score has the potential to improve radiologists' reading efficiency by **boosting their attention in the most suspicious examinations** while reassuring them in the less suspicious examinations. Furthermore, the research observed the learning curve of the radiologists, which suggests that more system practice could result in even faster reading times. As a secondary study endpoint, the AUC of the stand-alone computer system was compared to the AUC of the radiologists when examining mammograms in the unaided mode.

a.



b.

Fig. 21: (a) *Mammograms in a 71-year-old woman with invasive ductal carcinoma (outlined and with a computer-assigned level of suspicion score). When reading alone, four of 14 radiologists recalled the patient and 11 of 14 radiologists recalled the patient when using an Artificial Intelligence system for support. The outlined locations and scores are displayed in the AI system's viewer.* ***(b)*** *Mammograms in a 62-year-old lady without cancer who was recalled by 12 of 14 radiologists when reading unaided and by seven of 14 readers when utilizing an AI system. The outlined locations and scores are displayed in the AI system's viewer.*

As a result, the computer system's stand-alone performance was comparable to the radiologists' average performance. Although more research is needed to confirm these findings, this outcome suggests that **utilizing computers as a stand-alone first or second reader in screening programs may be viable**. Given the growing shortage of (experienced) breast radiologists, this could even

enable the development or continuity of screening programs. The findings of this research hence reinforce the trend that AI algorithms are approaching radiologists' performance in mammography breast cancer detection.

In conclusion, using an AI computer system for support increased radiologists' diagnostic performance in the detection of breast cancer during mammography without requiring additional reading time. However, as encouraging as these findings are, more research in a screening setting is needed to confirm them and to determine the true impact of AI support.

2.2 Negative aspects of AI in Medicine: ethical issues and unexpected consequences

Although the application of AI/ML in medicine has already yielded numerous beneficial results, as shown in the preceding section, there are still some ambiguities surrounding these revolutionary technologies to support healthcare operations.

One of the most pressing concerns is "**over-dependence**", which refers to physicians' growing reliance on automation's skills at the price of their own competence. The result of this overreliance on AI/ML decision assistance can lead to the *deskilling* of medical staff over time. The phenomenon of deskilling would become more evident when the technology failed or ceased to function. In an analysis conducted by a group of researchers from the City University of London on the reading of 180 mammograms by 50 practitioners, a **14.5 percent reduction in diagnostic sensitivity** for breast cancer detection was documented in the more experienced practitioners when they were presented with difficult-to-read images accompanied by computer interpretation, while only a 1.6 percent increase in diagnostic sensitivity was documented in the less experienced practitioners. These findings highlight not only how practitioners' overreliance on Machine Learning systems can affect their performance, but also how much more research is needed to understand the dynamics of this phenomenon, particularly in relation to the different levels of experience among the physicians involved, and the varying difficulty of the cases they are presented with.

Another critical aspect to consider in the context of introducing AI to medicine is the potential for a **progressive underestimation of the context**, which is often difficult to represent and express explicitly, as opposed to what is more easily codifiable and expressible in words or numbers, i.e., data, which is required for any AI to work and be useful. AI systems that are fed by a finite number of discrete data and unable to incorporate elements that are not very or not at all "data-able," such as cultural, social, or psychological aspects of a patient, or organizational aspects of a hospital setting,

may lead physicians to overlook the function of these elements, which are essential for accurate and efficient management of the care pathway and irreplaceable pivots of the unique physician-patient relationship. Furthermore, the data submitted to the system during the training phase of the algorithm is chosen by physicians, thus it has already been considered correct and classified. This strategy has two major drawbacks:

1. The potential for a **disparity in image quality** between those images utilized in the training phase and those in the medical records. The photos utilized in the training phase have a high quality since they have gone through a cleaning and pre-processing process that would be impossible to maintain in everyday clinical practice. The quality of the so-called "real-world data" rarely matches the "ideal data" assumptions that feed a Machine Learning algorithm.

2. The possibility of **many "gold standards"** (test samples used as a reference to evaluate other tests) in the same clinical case, each yielding different results.

A recent study by Dharmarajan et al., concentrating on older patients hospitalized for acute cardiopulmonary disease who were diagnosed with heart failure, chronic obstructive pulmonary disease or pneumonia upon admission to the hospital, may be an example. Patients received regular medical treatment for two or more of the aforesaid diseases simultaneously during their hospital stay, and not just for the primary diagnosis made at the time of admission. Clinical images are frequently seen in "gray zones" in real-life clinical practice and are **difficult to link with "golden" diagnostic criteria** as given in medical texts or recommendations.

Another risk is related to the problem of assigning **medico-legal responsibilities** in the event that a physician chooses to employ ML systems for support which may lead to errors, or the attribution of duty in the event that the physician chooses not to follow the systems' suggestions. The introduction of new diagnostic or therapeutic tools that were not sufficiently supported by scientific certainty has led to the withdrawal of these tools on several occasions, owing to their ineffectiveness or even harm, as evidenced by the unfortunate late completion of well-conducted clinical studies.

Artificial intelligence may offer a unique chance to relieve overworked healthcare personnel, but it also has the potential to negatively harm the healthcare workforce. It might put pressure on healthcare employees' skills and require them to retrain to adapt to the usage of the new IT systems. For millennia, medical practice has been dependent on the human contact between practitioners and patients, thus new technologies must be introduced with caution to avoid jeopardizing these relationships.

A further extremely sensitive issue that has long occupied the attention of the Community and national legislators is **how to handle personal data**. In fact, despite the enormous benefits associated with the ability to quickly manage enormous amounts of aggregated data, digitization has enabled the creation of extensive databases to which an increasing number of subjects have access: as a result, we have seen an exponential increase in the risks associated with the processing of the aforementioned data, their illicit dissemination, and the potential for harming the dignity and fundamental freedom of individuals. The health sector is a prominent example of community sensitivity to the need for regulation, but on deeper examination, this is an incredibly crucial issue for most domains of AI application, which now encompass practically the entirety of an individual's daily environment.

There is ongoing discussion on whether AI "fits within existing legal categories" or whether a new category with its own unique characteristics and implications should be created. Although the use of AI in clinical practice has immense potential to improve healthcare, four fundamental ethical challenges must be addressed before AI can fully realize its potential: **(1) informed consent to utilize data, (2) safety and openness, (3) algorithmic fairness and biases,** and **(4) data privacy.** As Tedros Adhanom Ghebreyesus, director general of the World Health Organization stated: *"Artificial Intelligence, like all technologies, has immense potential to enhance the health of millions of people around the world, but if applied poorly, it can cause harm."* In light of this discussion, WHO has created six guidelines that individuals involved in the design, development, and implementation of AI systems should keep in mind in order for them to work in all nations.

1. Safeguard human autonomy

Humans must maintain control over health-care systems and medical decisions; patients' privacy and confidentiality must be preserved and protected through suitable legislative frameworks for data protection.

2. Promote people's welfare and safety, as well as the general good

By use case, AI designers must meet regulatory criteria for safety, accuracy, and efficacy, and provide well-defined guidelines. They must make quality control and improvement when using AI technologies.

3. Ensure readability and transparency

Prior to the design or implementation of an AI technology, transparency involves allowing documents and information to be published. To enable the right and ethical use of technology, this information must be easily accessible and searchable.

4. Encourage responsibility

Even though AI technologies execute certain jobs, it is the responsibility of the stakeholders to guarantee that they are deployed in the right circumstances and by adequately qualified personnel.

5. Ensure equity and inclusivity

Artificial intelligence technology must be built in such a way that it may be used and accessed equally by all citizens of the world, regardless of age, gender, income, ethnicity, sexual orientation, and religion, while also protecting human rights.

6. Encourage AI that is responsive and long-lasting

The usage of AI should be transparently checked on a regular basis to ensure that responsiveness fulfills expectations and meets the requirements outlined so far. Furthermore, AI systems should be built to reduce environmental impact and improve energy efficiency. Governments and businesses must be equipped to deal with the repercussions of implementing such technology, such as educating healthcare staff to utilize AI systems.

To summarize, implementing AI in healthcare can improve access to treatments, particularly in terms of prevention, as well as citizens' or patients' quality of life, as long as the principles outlined above are followed and the use of technology is always accompanied by a sense of responsibility and respect for human rights.

Chapter 3

Trust in Artificial Intelligence for Medical Diagnosis in Italy: A survey study on Sicilian and Lombard medical staff

As described in the previous chapter, while AI/ML systems are gradually becoming more prevalent in the field of medicine, where the areas reliant on imaging will be among the first to be impacted by advances in AI technologies, there are still concerns and criticisms about their widespread adoption. The potential barrier to the effective implementation of AI/ML to support clinical decision-making stems from the nature of medical science, which is not deterministic and is always

subordinated to subjective interpretation and evaluation by the physicians, based on their own experience. As a result**,** the ultimate purpose of this thesis is to explore the perceptions of Italian healthcare personnel and students about the usage of Artificial Intelligence in medicine in Italy, as well as their level of understanding of the field. According to the DESI (The Digital Economy and Society Index), **Italy is the fourth-last European country for digitization** and the **last country for digital skills**. Thus, AI applications in Italy are still in their infancy when compared to other European states.

Up until now, the technical development of AI systems was the focus of attention; there is instead very little experience with the general public's attitude and the potential repercussions of incorporating AI systems into healthcare practice. Furthermore, there has been no research into the perspectives of healthcare professionals in Italy regarding the uses of AI programs in the medical area, and this must include future professionals, as students will be affected by the possible applications of AI in their employment. The goal of this study is to explore medical physicists' and students' awareness toward AI systems, as well as their perceptions and knowledge of the relevance and impact of them.

3.1 The Survey: Context, Objectives and Approach

Two separate surveys were conducted for the purpose of this research, one for medical workers and healthcare students that work or study in the region of Sicily, and one for medical workers and healthcare students that work or study in the region of Lombardy. This was done to investigate both perceptions on the employment of AI in the medical profession, and the differences in the opinions of the two populations.

As previously mentioned, the study's major goal is to evaluate medical personnel's and students' understanding and attitudes toward AI, the direction of AI growth in medicine in Italy, and the potential risks of adopting AI in the medical profession. Furthermore, the study intends to evaluate the primary benefits and drawbacks of deploying such systems in healthcare, as perceived by respondents, as well as how soon they believe it will be feasible to apply these technologies in their area. The questionnaire is broken into four sections, the first of which includes general information such as age, occupation, and place of employment/study. The second section focuses on the respondents' understanding of AI. The final segment discusses people's fears about AI's use. The fourth and last section focuses on respondents' readiness to incorporate AI into clinical practices. The survey was done in Italian for the sake of this thesis and translated into English afterwards.

Each participant responded to the survey through a link to the online questionnaire. The purpose of the survey was explained to participants in the questionnaire's prologue. Informed permission was implied by freely participating to the survey after receiving enough information about its purpose. The identity of the respondents is automatically anonymized by the Qualtrics platform, selected for the aim of this study. The Captcha method was employed within the questionnaire to guarantee that respondents were real people rather than programs designed to spam the survey, as well as to prevent and identify careless, rushed, or unplanned responses. Through this mechanism, the respondent was shown an image (or "challenge") of words or characters, and was required to input those characters accurately in order to proceed.

Survey

The survey was conducted from March to May 2022, consisting of 19 closed-ended and one open-ended questions. The answer to question 15 was assessed using a five-point ordinal scale, with 1 being the lowest and 5 the highest score. The following were the questions asked to the public:

Q1: How old are you?

a. Less than 25
b. 25-35
c. 35-45
d. 45-55
e. 55+

Q2: What is your occupation in the medical/healthcare field?

a. Physician
b. Dentist
c. Veterinary Physician
d. Nurse
e. Student
f. Other

Q3: In what province of Sicily/Lombardy do you work or study?

Q4: How long have you been practicing your profession? (question destined only to workers)

a. Less than 5 years
b. 5-10 years
c. 10-20 years

d. More than 20 years

KNOWLEDGE AND PERCEPTIONS RELATED TO ARTIFICIAL INTELLIGENCE

Q5: Artificial Intelligence is "the ability of a machine to exhibit human capabilities such as reasoning, learning, planning and creativity" [Europarl.europa.eu]. Are you aware of this?

a. Yes
b. No

Q6: How do you consider your knowledge related to Artificial Intelligence and its applications?

a. Very poor
b. Poor
c. Average
d. Good
e. Very good
f. Excellent

Q7: Currently, Artificial Intelligence has many applications in medicine (e.g., AI-assisted robotic surgery). How familiar are you with these applications?

a. Very unfamiliar
b. Unfamiliar
c. Moderately familiar
d. Very familiar
e. Extremely familiar

Q8: Have you ever used AI-based healthcare services? (Such as AI embedded in smart medical devices)?

a. Yes
b. No
c. I do not know

Q9: Do you agree that AI has useful applications in the medical field?

a. Strongly agree
b. Agree
c. Neither agree nor disagree

d. Disagree
e. Strongly disagree

Q10: Do you agree with the statement "the field of (your specialty, for example: radiology) will improve with the introduction of Artificial Intelligence"?

a. Strongly agree
b. Agree
c. Neither agree nor disagree
d. Disagree
e. Strongly disagree

Q11: "In my opinion, Artificial Intelligence will threaten/disrupt doctors' careers." Do you agree with this statement?

a. Strongly agree
b. Agree
c. Neither agree nor disagree
d. Disagree
e. Strongly disagree

Q12: In which field of medicine do you think AI is most useful?

a. Diagnosis
b. Deciding the correct treatment
c. Direct treatment (including surgery)
d. Biopharmaceutical research and development
e. Providing medical care
f. Other

Q13: What do you think is the greatest benefit of using Artificial Intelligence in medicine?

a. Faster health care
b. Improved patient access to disease screening
c. Reduction in the number of medical errors
d. More targeted medical care
e. Reduction in time spent by specialists on monotonous tasks
f. Greater uniformity in diagnosis and management decisions
g. AI has no emotional exhaustion or physical limitation

h. Better prediction of disease outcome
i. Other

CONCERNS RELATED TO THE USE OF ARTIFICIAL INTELLIGENCE

Q14: What do you think are the biggest drawbacks in the use of Artificial Intelligence in medicine? Select at most three.

a. Ceding health care to big data and technology companies
b. Not flexible enough to be applied to all patients
c. Medical liability affected by potential machine error
d. Concerns about data security and privacy
e. Decreased human interaction between patient and physician
f. Other

Q15: On a scale from 1 to 5, how much trust would you place in Artificial Intelligence used in medical diagnosis?

Q16: Would your trust related to the usage of AI increase if you could obtain knowledge about how it works, that could be understood by anyone, so that the degree of uncertainty could be reduced?

a. Yes
b. No
c. I do not know

PREPAREDNESS FOR THE INTRODUCTION OF ARTIFICIAL INTELLIGENCE INTO CLINICAL PRACTICES

Q17: If Artificial Intelligence systems were now adopted within the facility where you work, how would you rate the level of staff preparedness for such an event?

a. Very unprepared
b. Unprepared
c. Moderately prepared
d. Very well prepared
e. Extremely prepared

Q18: "AI should be included in the course of study and/or retraining courses of health care professionals." Do you agree with this statement?

a. Yes
b. No

Q19: What do you think is the time required before Artificial Intelligence is employed and has a noticeable impact on (your specialty) within Sicilian/Lombard medical facilities?

a. 1-5 years
b. 5-10 years
c. 10-15 years
d. 15+ years
e. Never

Q20: What are the obstacles in the implementation of AI systems in the medical facilities of the Sicilian/Lombard region, today and in the near future in your opinion?

3.2 Results and Discussion

Results

The results were tabulated and analyzed using the Qualtrics platform. For convenience, the data in this discussion will be approximated by excess.

Respondents under the age of 25 were quite numerous in both surveys (52% in Sicily, 44% in Lombardy); AI appears to be a field that students are interested in because of the potential applications that may affect them in their employment. 48% of Sicilian respondents were at least 25 years old (15% were between the ages of 25 and 35, 6% were between the ages of 35 and 45, 12% were between the ages of 45 and 55, and 15% were over the age of 55) and thus of working age. In the survey of Lombard healthcare staff and students, 56 percent of respondents were of working age (25-35 years old were the 8%, 35-45 years old were 11%, 45-55 years old were 14% and 23% were above 55).

As shown in Figure 23 and Figure 24, among the Sicilian respondents 29% were physicians, 47% were students, 14% were nurses and the remaining part were veterinary physicians, dentists, psychologists, physiotherapists and pharmacists. Among the Lombard respondents, physicians made up a larger proportion accounting for more than 34%, but there were fewer students (40%). Most of the respondents from both regions worked in the major provinces (35% of Sicilian respondents worked in Palermo area and 44% of Lombards worked in Milan area).

In the Sicilian survey, 36% of the working group has been practicing for more than 20 years, 24% for 10-20 years and 33% has been practicing for less than 5 years. Among Lombard respondents, those who have been practicing for more than 20 years make up the majority (30%), followed by those who began practicing during the last 5 years (18%) and those who have been practicing for 5-10 years (16%).

Fig. 22: *Sicily – practicing years*

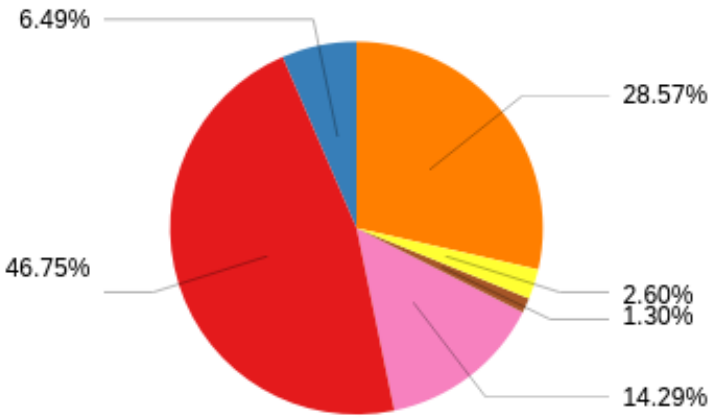
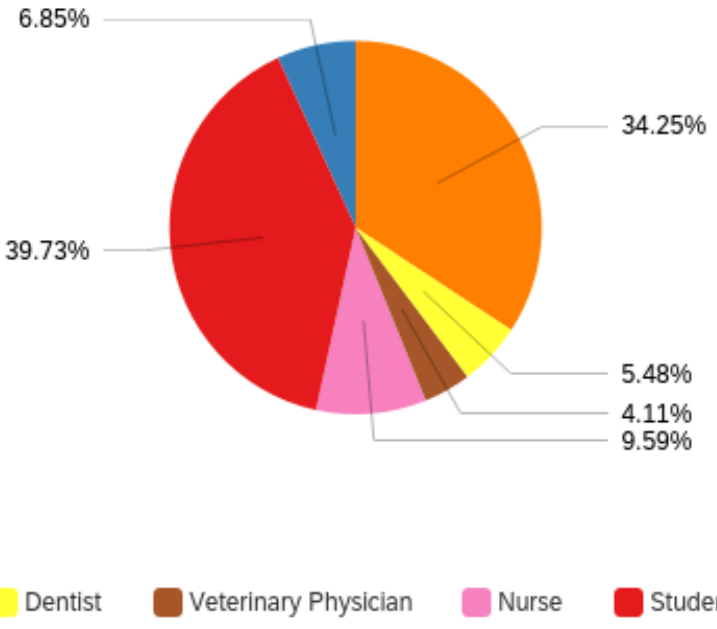


Fig. 23: *Lombardy – practicing years*



KNOWLEDGE AND PERCEPTIONS RELATED TO ARTIFICIAL INTELLIGENCE

As mentioned before, according to *Europarl.europa.eu*, Artificial Intelligence is *"the ability of a machine to demonstrate human qualities such as reasoning, learning, planning, and creativity"*; 97% of Lombard respondents indicated that they were aware. The scenario is different for Sicilian respondents, where 91% knew what Artificial Intelligence was and a remarkable 9% did not.

These insights also explain why, in the Sicilian case, 58% of respondents perceive their knowledge of AI and its general applications to be pretty limited (41% suggested “poor”, 17% claimed “very poor”). A remarkable 29% recognized themselves as "average" and 13% considered themselves above average (so, either good, very good, or excellent). Differently, the Lombard respondents have a higher opinion about their knowledge: a positive 30% perceives themselves above average (23% says "good", 5% suggests "very good", and 2% claims "excellent"), considerably higher than the Sicilian 13%. Furthermore, the majority considers themselves “average” (37%), while 33% claimed their knowledge was severely limited, quite below the Sicilian 58% [Figure 25, Figure 26].

Fig. 24: *Sicily – AI awareness*

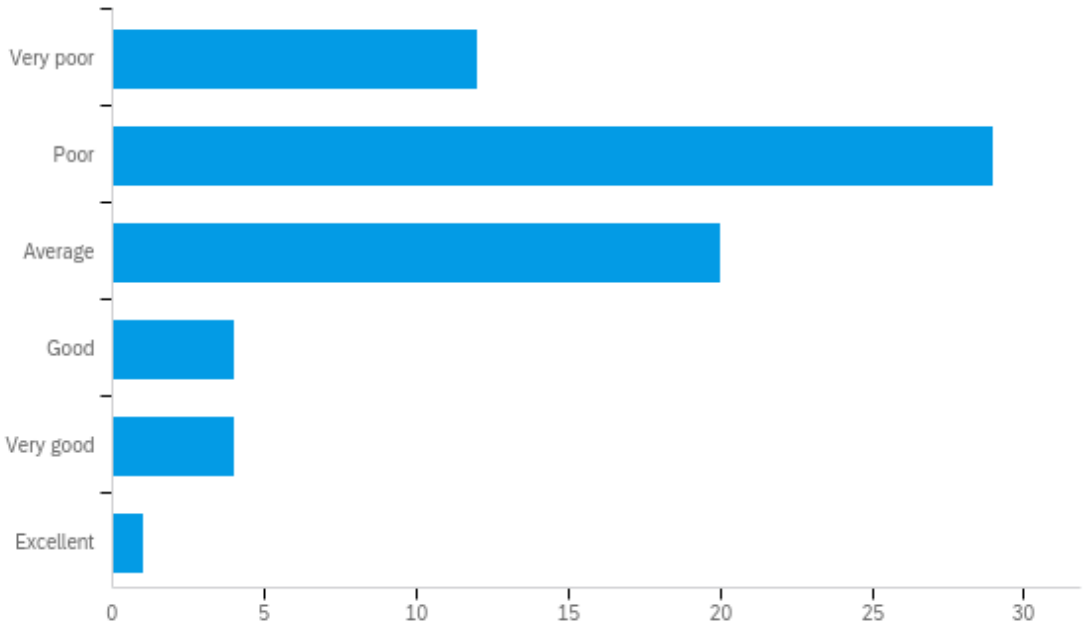
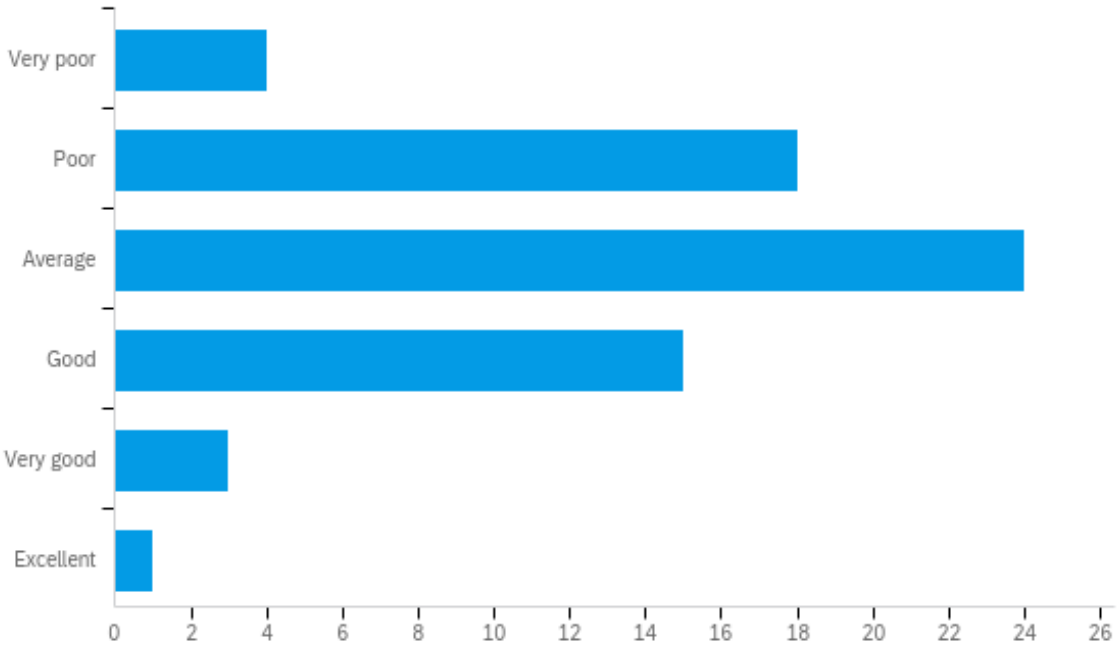


Fig. 25: *Lombardy – AI awareness*



Indagating the awareness of the respondents over AI applications in the medical field, such as **AI-assisted robotic surgery,** we can notice a slight difference in the familiarity with these technologies within the populations coming from the two different regions: in the case of Sicily, the majority is either unfamiliar or not familiar at all (51% unfamiliar, 30% very unfamiliar).

Fig. 26: *Sicily – awareness over AI medical applications*

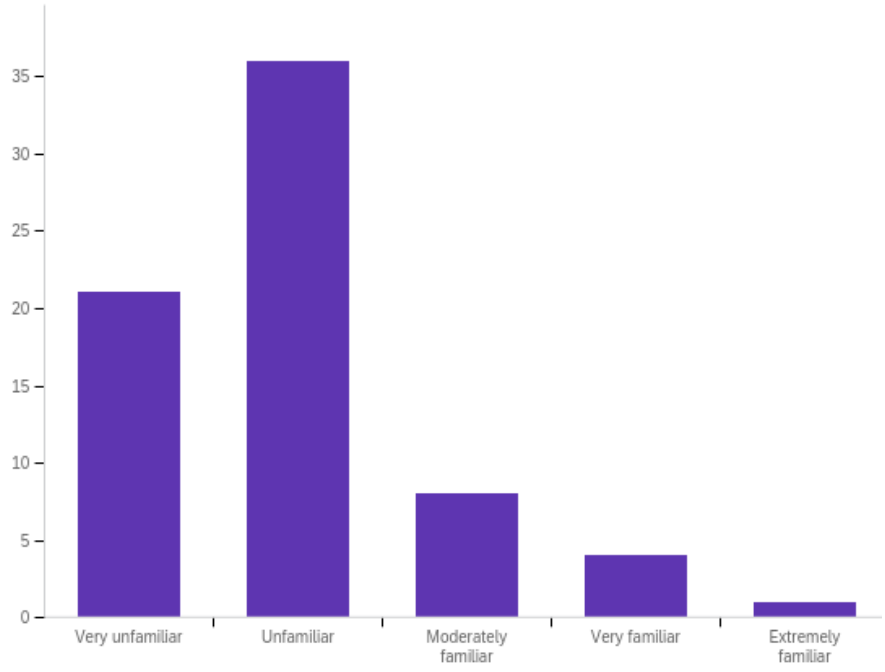
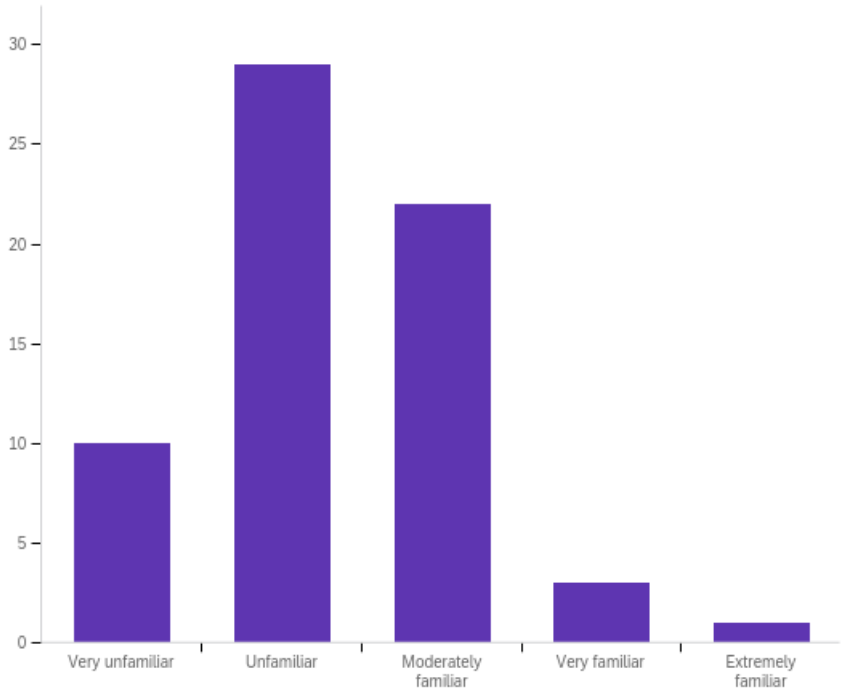


Fig. 27: *Lombardy - awareness over AI medical applications*



Only 18% is familiar with such concepts (11% moderately familiar, 6% very familiar and only 1% extremely familiar). In the case of Lombardy, the majority is either unfamiliar (45%) or moderately familiar (34%), and thus the people with an average or good degree of familiarity with AI in medicine are 40% (adding the 5% which is very familiar and the 1% which is extremely familiar). As a result, it appears that more people in Lombardy are acquainted with AI-based systems compared to Sicily [Figure 27, Figure 28].

In Sicily, 10% of respondents reported they have used an AI-based medical device at least once; in Lombardy the proportion reached 32%: a noticeable difference. Still, many people have never dealt with such technologies (76% for Sicily, 32% for Lombardy).

Even if a massive number of users have limited knowledge over AI uses in medicine, the majority in both surveys reports they do agree it has useful applications in the medical field (less than 2% disagrees in both cases). Similarly, in the surveys the largest percentage of participants (more than 75%) agrees that the **field of their specialty will improve with the introduction of Artificial Intelligence**. Moreover, the participants are not afraid of such technological changes: the largest part disagrees that AI would potentially threaten or disrupt doctors' career.

Participants in Sicily believe AI is most useful for diagnostic purposes (33%), followed by direct treatment including surgery (27%), and eventually for scientific research (26%). Only 1% believes that the most useful field is the identification of the correct treatment for the patient. The

majority of Lombard respondents (37%) believe AI applications are more effective for direct treatment rather than diagnosis, while a sizable proportion (31%) agrees with Sicilian respondents that diagnostic reasons are the most valuable for AI applications. Again, only a small proportion of individuals voted for "Deciding the correct treatment." [Figure 29, Figure 30].

Fig. 28: *Sicily – most important fields of applications of AI*

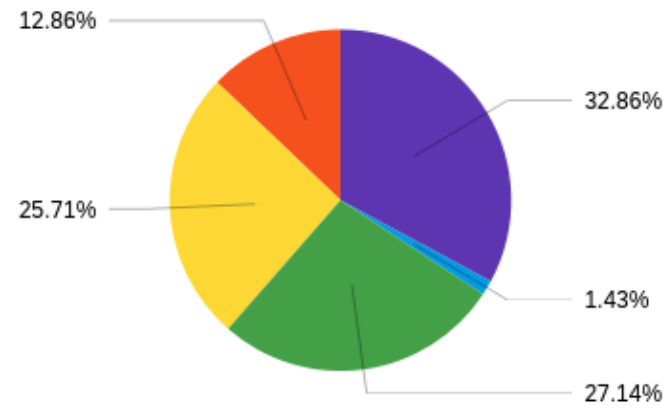
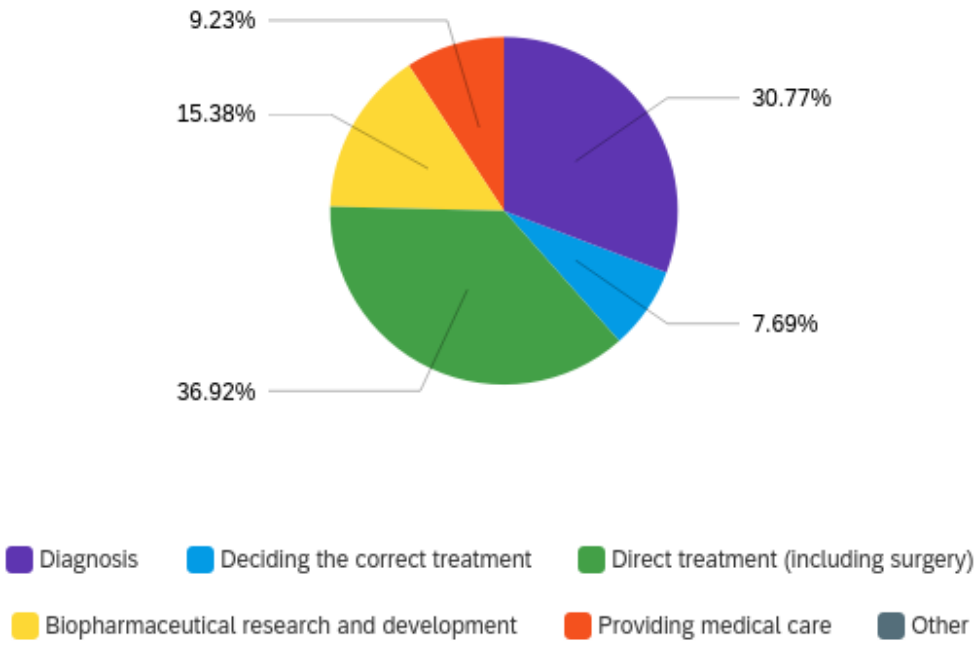


Fig. 29: *Lombardy – most important fields of applications of AI*



So, what are the real benefits of AI in medicine? Respondents agreed that AI will aid in the reduction of medical errors, faster healthcare, and more targeted medical care. In the Sicilian study, only a minority (6%) views uniformity in diagnostic and management decisions as the most significant benefit; in Lombardy, a minor fraction views AI's opportunity to prevent mental and physical tiredness as the most significant advantage [Fig. 31, Fig. 32].

Fig. 30: *Sicily – Benefits of AI in medicine*

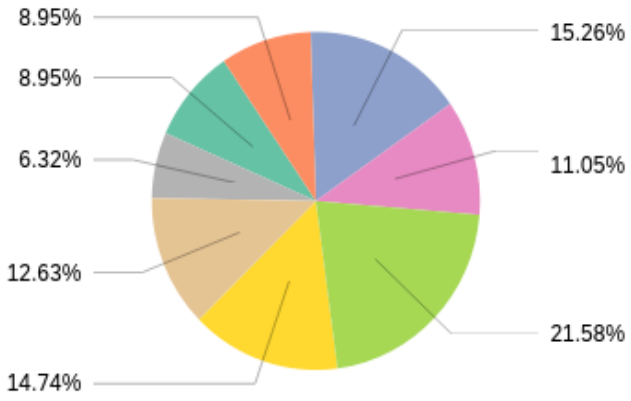
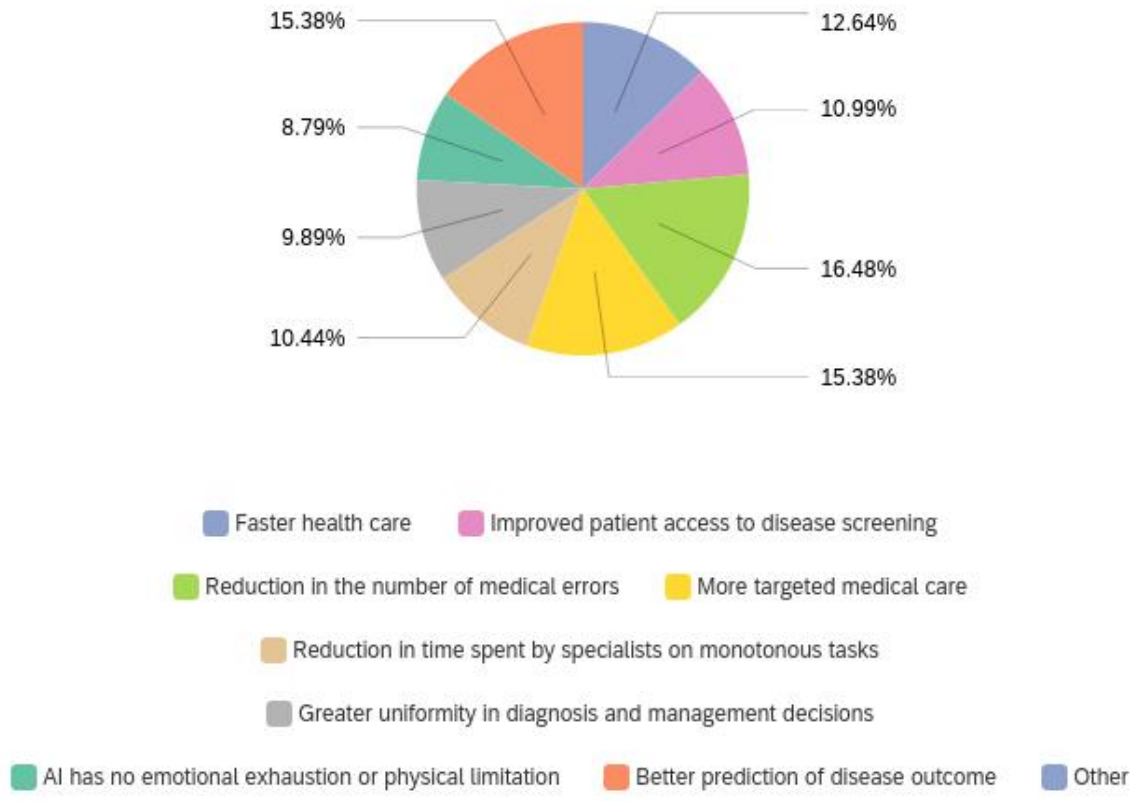


Fig. 31: *Lombardy – Benefits of AI in medicine*



CONCERNS RELATED TO THE USE OF ARTIFICIAL INTELLIGENCE

Participants in both studies agree that the major disadvantages and concerns associated with the implementation of AI technologies are the potential decreased human interaction between patients and healthcare staff, the possibility that the AI will not be flexible enough to be applied to everyone, and the medical liability for machine errors.

Fig. 32: *Sicily – AI disadvantages in medicine*

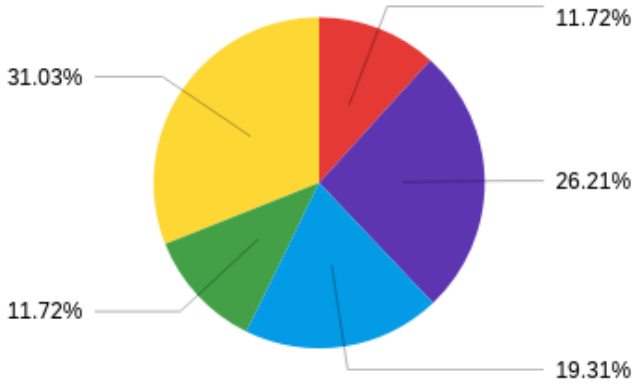
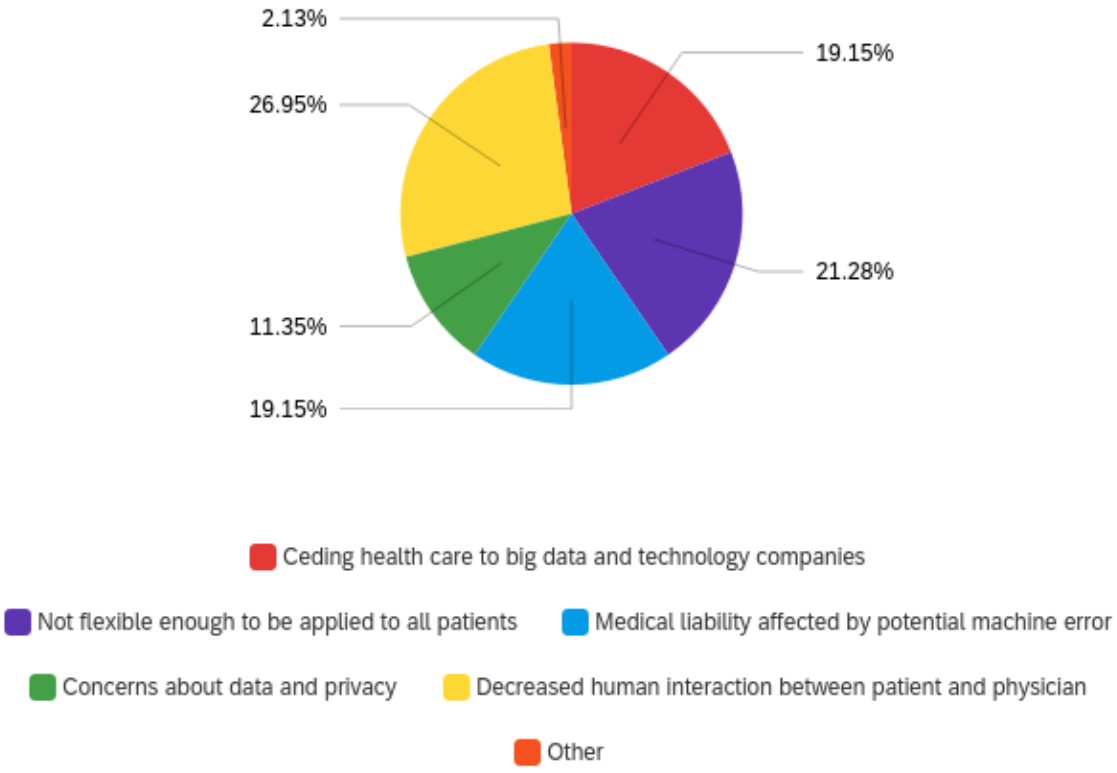


Fig. 33: *Lombardy – AI disadvantages in medicine*



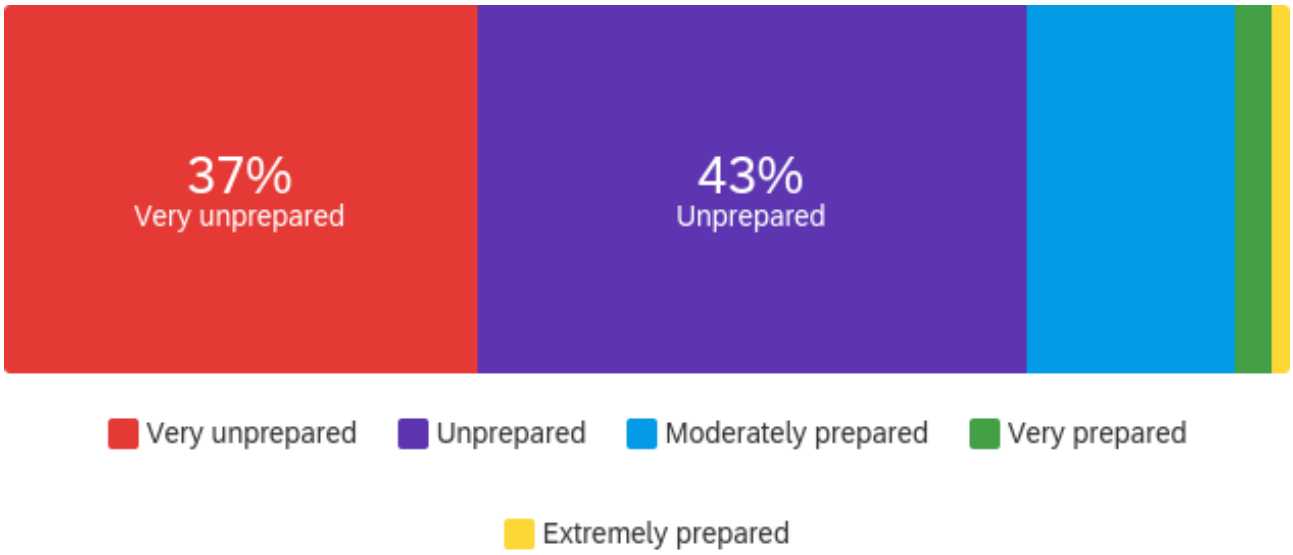
Instead, few people appear to be concerned about data and privacy. Some have responded "other," stating that it would be difficult for their colleagues to learn how to use these systems and evaluate AI's conclusions [Figure 33, Figure 34].

Furthermore, Lombards are more likely than Sicilians to trust the use of Artificial Intelligence in medical diagnosis, with a mean of 3.67 for the former and 3.17 for the latter on a scale from 1 to 5, with 1 being the lowest and 5 being the highest. Their trust would increase if they could obtain knowledge about the functioning of the AI system, in a way that it could be understandable by everyone, even by the ones not skilled in the art.

PREPAREDNESS FOR THE INTRODUCTION OF ARTIFICIAL INTELLIGENCE INTO CLINICAL PRACTICES

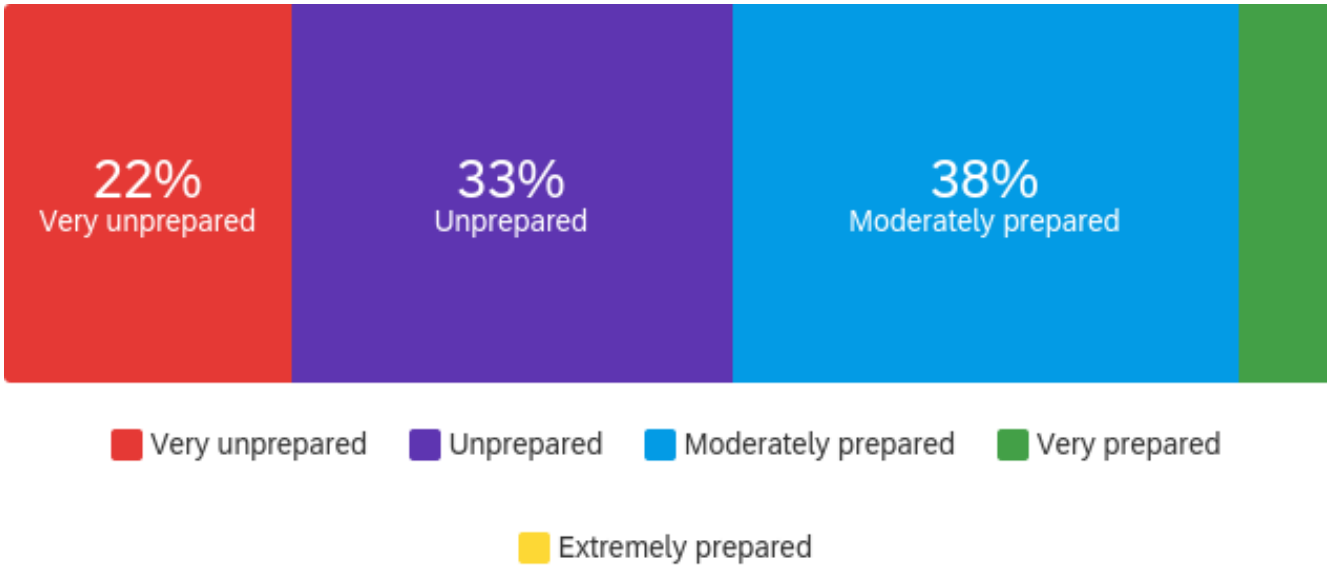
What would happen if AI was suddenly introduced in the facility where you work? 43% of Sicilian respondents think their colleagues would be unprepared for such change, while 37% claim they would be very unprepared [Fig. 35].

Fig. 34: *Sicily – preparedness for the introduction of AI*



Different scenario for Lombard participants: 33% considers their colleagues unprepared, 22% considers them very unprepared but a remarkable 38% values their colleagues as moderately prepared. Apparently, Lombard facilities appear to be more eager to implement new technologies [Figure 36]

Fig. 35: *Lombardy – preparedness for the introduction of AI*



The largest part agrees to include AI subjects in the course of study and/or retraining courses of healthcare professionals, with 97% agreeing in Lombardy and 93% in Sicily, where a small minority disagrees, explaining that "AI should not be used at all" or that a new figure would be required to act as an intermediary between machine and physician.

AI will be employed and have a noticeable impact on the respondents' specialty within the facilities where they work in 10-15 years, according to more than 32% of Sicilians and Lombards. More subjects from Sicily have answered "never" compared to Lombardy (6% vs 3%) as well as 15+ years (28% vs 13%). We can conclude that Sicilian participants have a more negative point of view towards AI implementations compared to their fellows from the North [Figure 37, Figure 38].

Fig. 36: *Sicily – Time required for AI implementation*

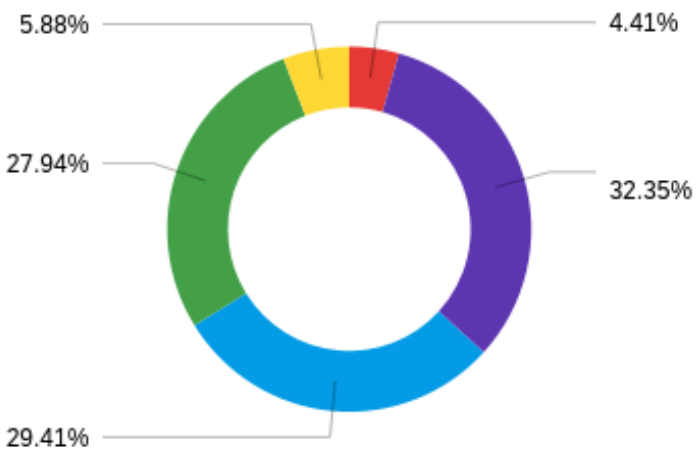
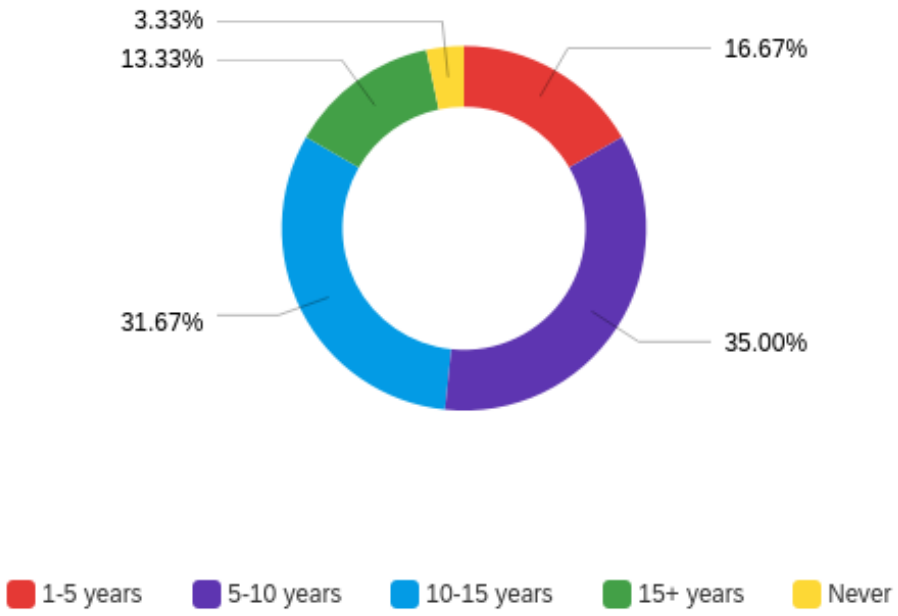


***Fig. 37**: Lombardy – Time required for AI implementation*



Discussion

To the best of my knowledge, this is the first study of Italian physicians and students' perspectives toward AI. According to the results of this survey, respondents are aware that AI is the ability of a machine to demonstrate human qualities such as reasoning, learning, planning, and creativity, but their knowledge of the practical applications of AI is limited, both in the general and in the medical field, where the majority has never worked with such technologies. They do, however, believe that AI is valuable in the medical sector and that it will improve the field of their specialty. AI, according to physicians and medical students, would be most useful for diagnosis and direct treatment, including surgery. The majority of Sicilian and Lombard doctors do not believe AI will replace them and do not perceive it as a danger or a threat to their career.

According to the study, most physicians and students agree that AI will help reducing medical errors and provide faster and more targeted medical care. Only a minority considers the potential of AI to provide greater uniformity in diagnosis and management decisions the most important benefit, as well as the ability to prevent emotional exhaustion or physical limitation, despite the fact that it is one of the most beneficial opportunities provided by the implementation of such technologies according to many studies.

The top three worries concerning the application of AI were (1) the potential for AI to reduce human interaction between patient and physician, (2) the likelihood of AI not being flexible enough to be applied to all patients, and (3) medical responsibility being impacted by potential machine error. Traditional physician-patient communications are being substantially altered by AI technology, and as a result, people may be concerned that they may lose face-to-face cues and personal encounters with clinicians. Thus, consumers may decline to adopt AI devices since they require human social engagement during service interactions. It is interesting to note how respondents appear unconcerned about data protection and privacy, even though health-related information is among the most sensitive pieces of information about a person. AI regulations and principles must be clearly defined today for the proper use of such technologies – having the potential to cause ethical and legal issues in medicine – and for increasing people awareness on the importance of their data.

The trust Lombards put on technology is higher compared to their compatriots from Sicily, maybe due to their slightly greater familiarity with AI applications. The overall trust of both populations would increase if they could obtain knowledge about the functioning of the AI-based system.

Sicilian healthcare workers believe their colleagues are unprepared for the quick introduction of AI technologies, whilst Lombard workers seem a step ahead again, and believe they are fairly equipped. Respondents from both regions agree that AI will be used and have a significant impact on their specialty in the facilities where they work within 5-10 years, and many believe AI should be part of the course of study of future medical personnel or of retraining courses.

Sicilian respondents think the obstacles in the implementation of AI systems in the medical facilities are:

- The facilities themselves, considered for the most part inadequate;
- The insufficiency of financial resources suitable for the purchase and maintenance of AI systems, or the incorrect use of them;
- The lack of technicians and medical personnel;
- Low flexibility of the healthcare staff, particularly of those with higher expertise who are reluctant to employ such computerized systems, together with their age which makes them less willing to embrace changes and innovation;
- The personnel made up for the majority by seniors, whose studies have been the most traditional ones, resulting in insufficient knowledge about AI-based applications;
- The insufficient preparation of recent graduates with respect to AI.

The Lombard counterpart has expressed similar opinions and introduced new ones. One of the respondents stated: “At present, funds for training and investment in new equipment follow other priorities given by the aftershocks of the pandemic. I am confident that once this period is finally over, the implementation of AIs will accelerate”. Many others have expressed dissatisfaction with:

- The Italian bureaucratic system, which is very slow;
- Institutional obstacles and the lack of a good regulatory system;
- Lack of funds and qualified staff, as well as training courses;
- The difficulty of convincing patients that an innovative treatment is effective and not dangerous, given their low trust on public healthcare;
- The modality of teaching all practitioners to use these systems, which might be challenging.

Conclusion

For many years now, the digital revolution has redefined people's lives and professional activities. The progressive integration of Artificial Intelligence into our daily life is one of the most innovative parts of this change, becoming established in all sectors; prominent among them is medicine, where great hope is placed in the contribution algorithms can make to healthcare.

The goal of this thesis was to analyze the implementation of Artificial Intelligence and Machine Learning algorithms in medicine and investigate the perceptions on AI of the Italian healthcare current and future staff. The aim is to highlight not only the positive results already achieved and the promising long-term prospects, but also the negative aspects and concerns associated with the introduction of these systems as decision-making support to medical personnel in the performance of their activities. The first section covered the history of Artificial Intelligence since its inception, introducing the fields of application of AI today and the most useful Machine Learning and Deep Learning algorithms that are beneficial for the medical field. The central part covered the applications of Artificial Intelligence in medical diagnosis, such as prostate cancer imaging, intelligent COVID-19 diagnosis, and breast cancer detection, leveraging their benefits.

In the first case study, it was demonstrated how Deep Learning algorithms can facilitate prostate cancer diagnosis by identifying the tumor area and by speeding up the "contouring" mechanism of radiologists thanks to the InnerEye software, issued by Microsoft. The study's findings suggest that by using the software alongside the physician's job, radiation planning can be completed 13 times faster, saving 90 percent of the doctors' time. As a result, professionals will have more time to spend with their patients for person-to-person interactions. Furthermore, it was demonstrated that

the AI's segmentation results were compatible with the contours outlined by the physicians, and their error stood within the permissible error bound.

In the second case study, I demonstrated the power of Artificial Intelligence in detecting and distinguishing COVID-19 pneumonia through the study of sounds, employing Convolutional Neural Networks to achieve near-perfect accuracy. This early diagnosis would allow for fewer hospitalizations and a higher chance of survival, while also providing a prospective smartphone application suitable for diagnosis to reduce the need for other cumbersome methods.

The use of Artificial Intelligence in the detection of breast cancer through mammograms was introduced in the third case study, showing that radiologists with AI support had better diagnostic performance than those who worked unaided, and that the use of these technologies dif not result in increased duration of the diagnostic process.

At the end of the second Chapter the negative aspects of Artificial Intelligence were discussed, as well as current critical concerns related to its regulatory framework, ethical issues and possible unintended consequences. Indeed, it should be emphasized that, while these technologies can make a significant contribution to medical activity, they cannot disregard the human impact of the physician, who must always make ultimate meaning of the examined data due to the non-deterministic nature of medical science. As a result, it becomes critical to regulate the use of Artificial Intelligence, particularly in such a delicate topic as medicine.

In the last section of this thesis, I introduced a study that I personally conducted and that, to the best of my knowledge, is the first research to evaluate the perceptions and awareness of current physicians and future healthcare professionals on the use of AI in their workplaces in Italy, with a major focus on Sicilian and Lombard personnel. The findings were intriguing and allow for the drawing of conclusions that might bring about a genuine shift in the Italian mode of operation, particularly in terms of technology – with Italy ranking last in Europe for digital capabilities – education and regulations.

The applicability of Artificial Intelligence in Italian medical facilities was analyzed within the limits of the survey. Following an evaluation, it was found that the medical professionals have limited knowledge of Artificial Intelligence, which must be improved. Even if the majority of participants in both surveys have a general idea of what Artificial Intelligence is, a significant 9% of the Sicilian population has never heard about it. This also implies that many doctors and students in Sicily are unaware of the most basic AI applications. Despite the enormous benefits represented by this technology, detailed in Chapter 2, less than 20% of Sicilians are familiar with Artificial Intelligence

applications in medicine. In contrast, 40% of Lombards are aware of these applications, with 32% having used AI-based medical equipment at least once, compared to 10% of Sicilians. Despite disparities in knowledge and perceptions, more than 98% recognized that AI is helpful in the medical sector, and more than 75% agreed that it will improve their field of specialization. The areas of medicine where respondents believed AI would be most effective were diagnosis and treatment planning.

Interestingly, despite their limited expertise, participants indicated that they are not scared by the adoption of these systems and did not consider AI as a threat that could entirely replace them in their profession. As a result, individuals are encouraged to adopt AI because they recognize it will never replace the physician's work, removing one more barrier to its eventual implementation. Indeed, if this fear was widespread, most people would be disinclined to support the deployment of Artificial Intelligence. Respondents' main concerns, on the other hand, include the fear of losing communication between doctor and patient, the liability issues in case of machine errors, and the flexibility of such technologies, which is viewed as not adaptive to all patients.

Alongside the first issue, it is true that AI technology might significantly alter the traditional physician-patient relation, hence customers may refuse to adopt AI devices due to their demand for human social contact during service encounters. On the contrary, we have seen for instance how the software InnerEye can help saving the time doctors devote to diagnosis, allowing them to spend more valuable time with the patient. Therefore, it is critical to examine all the trade-offs of any technology before implementing it, as well as to enhance the transparency of such applications for the benefit of patients and physicians.
Regarding transparency, it is evident from the study that when patients do not comprehend the inner workings of AI devices, they may exhibit lower trust in their operations. Because the nature of technological activities generates a lack of transparency, current AI systems used in healthcare are perceived by users as black boxes, which function as a barrier to AI technology adoption. Hence, individuals are less likely to employ AI-based tools in the future if the degree of uncertainty connected with their use is high. One possible solution is to provide explicit instructions about the inner workings of the machine that anyone who is not skilled in the art can understand, in order to boost trust in AI and eliminate ambiguity. In practice, trust grows as these technologies become more transparent.

According to the survey results, the second key concern is the accountability for AI-based decisions when errors arise while employing intelligent systems. There is currently no defined legal framework governing AI, which causes confusion in its proper usage and deployment, as well as

issues over who should bear accountability. Thus, this will generate a precarious situation for both professionals and patients if it is still unclear who will be held accountable if AI-based technologies provide incorrect suggestions in healthcare. The solution is for regulatory agencies to set normative criteria and assessment procedures for the adoption and use of AI in medicine, in conjunction with healthcare institutions. Regular audits and ongoing monitoring and reporting systems must be employed to continuously check the safety, quality, transparency, and ethical elements of AI-based services, and a clear policy on Artificial Intelligence is required prior to its adoption, while always respecting human rights and the WHO principles for the design, development, and implementation of AI systems.

Another major concern is that respondents are not aware of the importance of preserving their sensitive data for their protection and privacy: in healthcare services, respecting a patient's privacy is an essential ethical principle because it is associated with wellbeing and personal identity. Thus, healthcare professionals should respect patients' confidentiality by safeguarding their health information, avoiding secondary use of data and building a rigorous mechanism for obtaining informed consent from them. Patients will suffer psychological and reputational harm if their privacy needs are not satisfied. Users' personal data (such as habits, preferences, and health records) are likely to be preserved and shared across the AI network, jeopardizing their identity. For these technologies to be deployed, a new figure in charge of establishing AI systems in medical facilities is required: a role responsible for making technological decisions, appointed within the various Italian provinces.

Finally, the Italian government must invest more in technology and innovation and make better use of its funds, beginning with education and school systems, both in primary and secondary schools, as well as universities and post-graduate training or retraining courses: given the rapid digital revolution today's society is facing, it is vital to stay up to date by introducing technology-related subjects in the educational process of individuals and, in our case, of healthcare personnel. In fact, many people are deterred from using and experimenting AI devices because of disinformation generated by the traditional school system, which is resistant to change. Therefore, AI-related subjects and hands-on training should be included into the healthcare professionals' curriculum as soon as possible. Finally, it is critical to educate not only medical practitioners, but also consumers and patients, in order to increase their trust in technology and lay the groundwork for the rapid development of AI-based systems, which, as we have seen, create life-changing benefits that will lead to unprecedented innovations in the medical field.

Bibliography and Sitography

Wikimedia Foundation. *Intelligenza Artificiale*. Wikipedia. Retrieved March 13, 2022, from https://it.wikipedia.org/wiki/Intelligenza_artificiale

Wikimedia Foundation. *Computing machinery and intelligence*. Wikipedia. Retrieved March 13, 2022, from https://en.wikipedia.org/wiki/Computing_Machinery_and_Intelligence

DataScienceGyan. (2018, June 5). *Artificial Intelligence vs machine learning vs deep learning*. DATASCIENCEGYAN. Retrieved June 5, 2022, from https://datasciencegyan.com/artificial-intelligence-vs-machine-learning-vs-deep-learning/

La Storia dell'intelligenza artificiale, da turing ad Oggi. CyberLaws. (2018, November 21). Retrieved March 13, 2022, from https://www.cyberlaws.it/en/2018/la-storia-dellintelligenza-artificiale-da-turing-ad-oggi/

Gupta, N., and R. Mangla. *Artificial Intelligence Basics*. Mercury Learning and Information, 2020.

Russell, Stuart J., et al. *Artificial Intelligence: A Modern Approach*. Pearson, 2022.

Chiara Casse. *Intelligenza artificiale Nella Finanza: 3 possibili applicazioni*. Capterra. Retrieved March 18, 2022, from https://www.capterra.it/blog/2191/intelligenza-artificiale-finanza#:~:text=L'intelligenza%20artificiale%20(AI%2C,aziende%20gestiscono%20il%20proprio%20patrimonio

harshpreet0508, arvindpdmn. (2022, January 12). *Supervised vs unsupervised learning*. Devopedia. Retrieved June 5, 2022, from https://devopedia.org/supervised-vs-unsupervised-learning

Wikimedia Foundation. *Apprendimento Automatico*. Wikipedia. Retrieved March 18, 2022, from https://it.wikipedia.org/wiki/Apprendimento_automatico

Wikimedia Foundation. *Algoritmo*. Wikipedia. Retrieved March 18, 2022, from https://it.wikipedia.org/wiki/Algoritmo

Wikimedia Foundation. *Training e test set*. Wikipedia. Retrieved March 18, 2022, from https://it.wikipedia.org/wiki/Training_e_test_set

Casadei, C. (2019, November 14). *Support-Vector Machine*. maggiolidevelopers. Retrieved March 19, 2022, from https://www.developersmaggioli.it/blog/support-vector-machine/

Wikimedia Foundation. *Regressione lineare*. Wikipedia. Retrieved March 20, 2022, from https://it.wikipedia.org/wiki/Regressione_lineare#Regressione_lineare_multipla

Wikimedia Foundation. (2021, November 3). *Modello logit*. Wikipedia. Retrieved March 20, 2022, from https://it.wikipedia.org/wiki/Modello_logit

Dennis, J. (2020, October 29). *Using k nearest neighbors with base python.* Medium. Retrieved June 5, 2022, from https://jakedennis877.medium.com/using-k-nearest-neighbors-with-base-python-fcafae344278

Wikimedia Foundation. (2021, October 24). *Apprendimento non supervisionato*. Wikipedia. Retrieved March 22, 2022, from https://it.wikipedia.org/wiki/Apprendimento_non_supervisionato

Galarnyk, M. (2022, April 27). *Visualizing Decision Trees with python (scikit-learn, Graphviz, matplotlib)*. Medium. Retrieved June 5, 2022, from https://towardsdatascience.com/visualizing-decision-trees-with-python-scikit-learn-graphviz-matplotlib-1c50b4aa68dc

Angeloonline. (2020, April 17). *Che cos'è il clustering?* AppuntiSoftware.it. Retrieved June 5, 2022, from https://www.appuntisoftware.it/clustering/

Wikimedia Foundation. *Deep learning*. Wikipedia. Retrieved March 23, 2022, from https://en.wikipedia.org/wiki/Deep_learning

Convolutional Neural Network (o CNN) cos'è e come funziona. Neuragate. (2021, April 20). Retrieved March 23, 2022, from https://www.neuragate.it/intelligenza-artificiale/rete-neurale-convoluzionale-cose-e-come-funziona/

Jeremy Jordan. (2020, November 8). An overview of semantic image segmentation. Jeremy Jordan. Retrieved June 5, 2022, from https://www.jeremyjordan.me/semantic-segmentation/

Swapna. (2022, January 25). Convolutional Neural Network: Deep learning. Developers Breach. Retrieved June 5, 2022, from https://developersbreach.com/convolution-neural-network-deep-learning/

2017, P. on 19 S. (2022, January 6). *Artificial Intelligence as a music composer*. Globant Blog. Retrieved June 5, 2022, from https://stayrelevant.globant.com/en/artificial-intelligence-composing-original-music/

SuperDataScience. (n.d.). Retrieved June 5, 2022, from https://www.superdatascience.com/blogs/convolutional-neural-networks-cnn-step-3-flattening

Kumar, P. (2021, August 25). *Max pooling, why use it and its advantages.* Medium. Retrieved June 5, 2022, from https://medium.com/geekculture/max-pooling-why-use-it-and-its-advantages-5807a0190459

What is feature detector and feature map in CNN? Quora. (n.d.). Retrieved June 5, 2022, from https://pythonandmlbasics.quora.com/What-is-feature-detector-and-Feature-Map-in-CNN

Maruzzella, di G. *Intelligenza artificiale e medicina: 3 Casi*. Indigo.ai. Retrieved April 7, 2022, from https://blog.indigo.ai/it/intelligenza-artificiale-medicina-tre-casi-di-successo/

Pictet Asset Management. *Intelligenza artificiale in Campo Medico: Quali Sono Le Applicazioni Oggi - Pictet per te*. Pictet Asset Management. Retrieved April 7, 2022, from https://www.am.pictet/it/blog/articoli/tecnologia-e-innovazione/intelligenza-artificiale-in-campo-medico-quali-sono-le-applicazioni-oggi#:~:text=I%20medici%20usano%20microscopi%20dotati,possibile%20con%20la%20scansione%20manuale.&text=Ma%20l'intelligenza%20artificiale%20pu%C3%B2,rendere%20pi%C3%B9%20efficiente%20la%20sanit%C3%A0

Oktay, O., Schwaighofer, A., Carter, D., Bristow, M., Alvarez-Valle, J., & Nori, A. (2020, November 30). *%*. Microsoft Research. Retrieved April 7, 2022, from https://www.microsoft.com/en-us/research/blog/project-innereye-evaluation-shows-how-ai-can-augment-and-accelerate-clinicians-ability-to-perform-radiotherapy-planning-13-times-faster/

Oktay, Ozan et al. "Evaluation of Deep Learning to Augment Image-Guided Radiotherapy for Head and Neck and Prostate Cancers." *JAMA network open* vol. 3,11 e2027426. 2 Nov. 2020, doi:10.1001/jamanetworkopen.2020.27426.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention*. Springer, Cham, 2015.

"A Microsoft AI Tool Is Helping to Speed up Cancer Treatment – and Addenbrooke's Will Be the First Hospital in the World to Use It." *Microsoft News Centre UK*. (9 Dec. 2020). Retrieved Aril 16, 2022, from https://news.microsoft.com/en-gb/2020/12/09/a-microsoft-ai-tool-is-helping-to-speed-up-cancer-treatment-and-addenbrookes-will-be-the-first-hospital-in-the-world-to-use-it/

Intelligenza artificiale e diagnosi Precoce del Covid-19 - mise. Retrieved April 22, 2022, from https://atc.mise.gov.it/images/documenti/Rivista/2020/Intelligenza_Artificiale_e_diagnosi_precoce_del_COVID-19.pdf

Wikimedia Foundation. (2019, June 2). *Spettrogramma*. Wikipedia. Retrieved April 22, 2022, from https://it.wikipedia.org/wiki/Spettrogramma

Wikimedia Foundation. *Continuous wavelet transform*. Wikipedia. Retrieved April 22, 2022, from https://en.wikipedia.org/wiki/Continuous_wavelet_transform

Rodríguez-Ruiz, Alejandro, et al. "Detection of breast cancer with mammography: effect of an artificial intelligence support system." *Radiology* 290.2 (2019): 305-314.

Hupse R, Samulski M, Lobbes MB, et al. Computer-aided detection of masses at mammography: interactive decision support versus prompts. Radiology 2013;266(1):123–129.

Intelligenza artificiale in Sanità Rischi e Vantaggi dell'utilizzo della Tecnologia. BollinoIT. (2021, July 12). Retrieved May 1, 2022, from https://www.bollino.com/2021/07/12/intelligenza-artificiale-in-sanita-vantaggi-e-rischi/

The Digital Economy and Society index (DESI). Shaping Europe's digital future. Retrieved May 1, 2022, from https://digital-strategy.ec.europa.eu/en/policies/desi

Caltagirone, I. J. (2022, March 13). *Survey questionnaire to investigate the knowledge and perceptions of Sicilian healthcare staff and students towards Artificial Intelligence (AI)*. qfreeaccountssjc1.az1.qualtrics.com. Retrieved May 31, 2022, from https://qfreeaccountssjc1.az1.qualtrics.com/jfe/form/SV_6rH4Q0wrpuAXBxI

Caltagirone, I. J. (2022, March 13). *Survey questionnaire to investigate the knowledge and perceptions of Lombard healthcare staff and students towards Artificial Intelligence (AI).* qfreeaccountssjc1.az1.qualtrics.com. Retrieved May 31, 2022, from https://qfreeaccountssjc1.az1.qualtrics.com/jfe/form/SV_50i6VOk3Xvl93vM