UNIVERSITA' DEGLI STUDI DI NAPOLI
"FEDERICO II"

DIPARTIMENTO DI SCIENZE BIOMEDICHE AVANZATE

TESI DI DOTTORATO
IN
SCIENZE BIOMORFOLOGICHE E CHIRURGICHE

DESIGN, IMPLEMENTATION AND REALIZATION OF AN INTEGRATED PLATFORM DEDICATED TO E-PUBLIC HEALTH, FOR ANALYSING HEALTH DATA AND SUPPORTING THE MANAGEMENT CONTROL IN HEALTHCARE COMPANIES

Supervisore Scientifico
Ch.mo Prof. Mario Cesarelli

Dottorando
Dott. Carlo Ricciardi

Supervisore Aziendale
Dott. Pasquale Russo

ANNO ACCADEMICO 2020/2021

A Debora e

Aurora

Summary

Abstract

In healthcare, the information is a fundamental aspect and the human body is the major source of every kind of data: the challenge is to benefit from this huge amount of unstructured data by applying technologic solutions, called Big Data Analysis, that allows the management of data and the extraction of information through informatic systems. This thesis aims to introduce a technologic solution made up of two open source platforms: Power BI and Knime Analytics Platform. First, the importance, the role and the processes of business intelligence and machine learning in healthcare will be discussed; secondly, the platforms will be described, particularly enhancing their feasibility and capacities. Then, the clinical specialties, where they have been applied, will be shown by highlighting the international literature that have been produced: neurology, cardiology, oncology, fetal-monitoring and others. An application in the current pandemic situation due to SARS-CoV-2 will be described by using more than 50000 records: a cascade of 3 platforms helping health facilities to deal with the current worldwide pandemic.

Finally, the advantages, the disadvantages, the limitations and the future developments in this framework will be discussed while the architectural technologic solution containing a data warehouse, a platform to collect data, two platforms to analyse health and management data and the possible applications will be shown.

1. Introduction

In the human activities the information is handled in different ways: informal ideas, natural language (spoken or written, formal or colloquial), drawing, graphs, schemes, etc. In the informative systems, the information is represented through data; thus, it is essential to design a definition for data and information in order to distinguish them. On the one hand, the information results from operations such as extraction and elaboration on data (they have different meaning on the basis of the context). On the other hand, the data is what we have immediately available before every possible processing operation.

In computer science, the data is an informative element, made up of symbols that need to be elaborated. It's clear that, without an interpretation, the data is not so much useful and, when correctly interpreted and correlated, the data has can provide researchers with information that can enrich everyone knowledge.

In the healthcare context, the information is the most important aspect and the human body consists in the major source of every kind of data: the challenge is to benefit from this huge amount of unstructured data by applying technologic solutions, called Big Data Analysis, that allows the management of data through informatic systems. Health data can be found in several forms: they have been recorded for long time on papers and it's still happening in small towns. Thanks to the introduction of digital format, the clinical data have been transformed in new formats. Health records include information in heterogenous formats: audio recording, computer tomography, magnetic resonance and other medical imaging, electrocardiogram, etc. The business intelligence (BI), also called clinical intelligence in healthcare, has a fundamental role for the development of medical processes including computerized methods and the procedures for extracting and transforming raw data in useful clinical information.

Data transformation is performed through statistical methods and by analysing unstructured and complex data. Similarly, machine learning and deep learning have transformed the medical framework. Indeed, technologies based on artificial intelligence has the potential to influence deeply social service and the whole healthcare context.

Data should be considered like a heritage and the attention paid to Big Data is testified by the will of health facilities to spend money for a better management of data and their policies in order to enhance what clinicians usually collect in their clinical activities.

Computer-based systems need to be developed as user-friendly in order to allow non-technical professionals to exploit the tools implemented by the information and communication technology (ICT).

This thesis aims to introduce a technologic solution made up of two open source platforms. Power BI and Knime Analytics Platform, that have shown their feasibility to produce interactive dashboards to support health management and to analyse data through machine learning by designing workflows containing the implementation of algorithms.

First, the importance and the role of BI and machine learning in healthcare will be discussed; the platforms will be discussed and described, particularly enhancing their feasibility and capacities. Then, the clinical specialties where they have been applied, will be shown by highlighting the international literature that have been produced. An application in the current pandemic situation due to SARS-CoV-2 will be shown: a cascade of 3 platforms helping health facilities to deal with the current worldwide pandemic.

Finally, the advantages, the disadvantages, the limitations and the future developments in this framework will be discussed.

1.1. The BI in healthcare

The BI tools have been developed to help the companies to exploit their information asset in light of strategic and tactical decisions.

BI allows to collect and handle the data in order to transform them in information useful for decision-making processes [1]. Its scope is to:

- Go from ideas to facts;
- Increase the amount and the quality of information in order to make it more significant;
- Spread and share information within a company.

The BI is a system made up of models, methods, processes, people and tools that make the regular and structured collection of data possible in a company. Moreover, elaborations, analysis and aggregation allow to transform data into information, to keep them, to make them available to the main actors of processes and to show them in an easy, flexible and efficacious form that effectively help in decision-making [2]. The BI systems, thus, involve:

- The collection of the data in a company;
- The phase of cleansing, validation and integration;
- The following elaborations, aggregations and analysis;
- The fundamental use of this big amount of data to create information and use it in the decision-making process.

These systems, considered for helping in the decision-making, are called Decision Support Systems (DSS), although the terminology is susceptible to changes due to the new functions that have been introduced in the framework.

There two types of' DSS":

- Data-Oriented, which consists in complex techniques to query a Data Warehouse (DW). There are two subtypes:
    - *Data Retrieval*: queries on several dimensions by using heterogenous sources.

    - *"Data Mining*" techniques: extraction of information and hidden pattern in large database thorough mathematical and statistical techniques.
- Model-Oriented, the context where the decision-making context happen is reproduced together with its possible effects. The existing models are:
    - *Predictive*,
    - *"What if*,
    - *Phi optimization.*

The context, where the organizations, the companies and the resources (both human and technologic) work, has determined the grade of opening towards the innovation, thus an evolutional process of informatic systems. The first level consists in creating a database allowing to collect large datasets which the companies interact daily with.

The next step is to make the old data available to the policy makers which are part of the decision-making process; that's why the development of a DW is necessary, just like a container for all the data structured in a way that could be helpful in the organizational processes.

The structured disposition of data allows BI tools to analyse the performance of the company, the make forecasting on the future performance and to show these analyses to the direction that will use them to take key decisions.

In Italy, the health companies and the hospitals can acquire tools for managing their data both in a managerial and a clinical side. There are several examples in literatures of BI applied in healthcare [3-6].

1.2. Machine learning in Healthcare

In literature, data mining is a recent science, that has been developed by taking a cue from other science, i.e. informatics, marketing, statistics. In particular, several methodologies, used in data mining, were born from two communities: the former is informatics, related to machine learning, and the latter is statistics that were involved in the analysis of data and computational statistics.

The novelty offered by data mining is the integration of previous methodologies with more practical aspects. Despite being different based on the applicational context, the aim of data mining was to produce interpretable and useful results for supporting business decisions. Statisticians have always dealt with building methods and models for analysing data, although usually only theoretically. Some of them have paid attention to the practical aspects and, starting from Eighties, due to the rising importance of the computational side, they developed new researches in this context. This situation has led to the implementation of the above-mentioned computational side starting from Nineties. An example is represented by the bootstrap, the Monte Carlo method, the Markov chain and probability-based systems [7-9].

Statisticians at the same time were interested in machine learning. It is the last piece of the story for the birth of data mining in the second part of Nineties. After the spread of the informatic methods, there was an increasing amount of available data and information, often in non-traditional formats: this was due to browsing the internet, text and multimedia data.

Innovative and original researches were born in this situation; they were not part of a specific discipline, but they were surely part of data mining research.

To sum up, machine learning is a branch of artificial intelligence that uses statistical methods to implement algorithms, capable of recognizing hidden patterns among data and making prediction on them by building models data based.

It is possible to identify some applicational scopes for which specific language is require. They are listed here [10]:

- Scoring system: this is an analytical approach based on assigning to each client (prospect) the probability of adhering to a marketing campaign. The scope is to classify client or eventual prospects and implementing marketing strategies that are based on specific targets by building a predictive model; it should be capable of identifying relationships among behavioural and target variables. The output of these models is a score with the probability of having a good outcome with the marketing campaign;
- Credit scoring: it is a particular case of scoring system which assesses each client based on some variables that describe his behaviour as regards payment. A numeric score is computed, and it represents how much the client is a good payer. This analysis can be used to understand whether it is possible to provide financial credit by using the risk class of the client;
- Customer profiling: it is an application of clustering techniques to detect homogeneous groups based on behavioural and demographical variables. Detecting the different types of customer allows to conduct strategic and customer care marketing campaigns. The present and future value of the client (by assigning him in a class) can be determined in order to allocate him in groups of customer service, priority of purchases or delay in payment;
- Market basket analysis: association techniques are used to detect which products are acquired together by clients. It is useful to allocate products in a market and to encourage the customers to buy more things, but also to make more efficacious marketing and merchandising strategies;

- Fraud detection: it is based on creating personalize profiles to evaluate the inclination of new clients to commit a fraud when making new subscriptions;
- Claims settlement: an insurance can be interested in investigating some accidents in order to understand which factor can reduce the time to claim settlement. It consists in finding the data, the behaviours, the situations that can be considered abnormal while the aim is to reduce everything is considered not optimized, excessive or a mistake;
- Text mining: it is the application of data mining to documental data, text files, articles, memos, clinical records, patents, reports, questionnaires, e-mail, etc. this application often consists in using clustering to identify homogeneous groups of words as regard a topic; it makes easier the capacity to find a theme and identify relationships among topics. Some of these techniques can be applied online and are called Web Mining. Here is a list of examples:
    - Click-stream analysis: it is the analysis of the behaviours of people visiting a website in order to understand which pages determines the online purchase a product. These techniques support the design and the implementation of online advertising campaigns to reduce costs and make the gain rise.
    - Dynamic contests targeting: it consists in showing dynamically the content of a page based on the profile of the visitor.

Data mining represents a fundamental resource also in the healthcare sector because of the big amount of data that is produced daily, they are:

- Related to patients, hospital stay, rehabilitation and health status;
- Obtained through the modern and advanced imaging diagnostics devices;
- Related to logistics and management of health facilities.

All these data are often unexplored and machine learning makes them a valuable starting point for studying correlations, relationships and unknown phenomena.

As a result, it becomes possible developing algorithms that allows to obtain outcomes supporting clinical decision-making, both in early diagnosis and in choosing therapies and rehabilitation strategies.

Finally, by analysing the results provided by these techniques, it is possible to better understand risk factors, causes and mechanisms that conduct to some pathologies and to identify common pattern among all patients [11].

1.3. Control management in health companies

Starting from Eighties, a greater awareness has risen regarding the necessity to implement managerial tools which allows to make efficient and efficacious the methodologies utilized for health services. Management control is needed in health companies because it represents the "analysis, evaluations, decisions and actions" that are useful to improve the interplay among technical, professional and economic aspects, despite being conscious that public health companies represent a public service (no-profit) with a social scope (to safeguard the health of the population).

In Italy, the hospital became companies through a process based on economic and financial analysis, design and managerial tools and the development of new systems for accountability. The strategical planning and programming, supported by the organization of the company, allow to manage the activities, assess and evaluate them according to efficacy and efficiency and, eventually, modify them in order to reach a specific goal.

Management control is a short-term process through which health policy makes sure that the resources acquired are used efficiently in order to achieve the objectives set out in strategic planning and programming.

It is an internal process for management and healthcare professionals, and it allows managerial accountability; it is also a guiding tool for the operators who receive some objectives and can understand whether they are satisfying the requests they received. The programming and control systems also serve as a communication tool for the workers in order to stimulate them to achieve some fixed objectives.

The healthcare sector is undoubtedly characterized by a very high innovative content but, in the majority of cases, new technologies determine an innovation in the products: today it is possible to deal with diseases that were not curable years ago, an extraordinary progress has been made not only in the therapeutic field but also in the prophylactic, rehabilitative and diagnostic ones.

Healthcare spending have increased more and more over the years, due to both the demand (the improvement in the standard of living, the aging of the population, health intended as physical, mental and social well-being...) and supplied service (increase amount of doctors and costs).

The corporatization of health facilities has imposed a gradual transition to an economic-asset accounting, made up of the set of surveys which allow to identify the costs and revenues connected to market and to follow the financial movements of management.

The importance of management control in the health sector, in addition to accounting control, derives from the fact that the management accounting takes into consideration the results of the company as a whole while cost accounting and other management control tools look at the results of specific parts of the company, which allow to identify the various issues, to analyse the particular economic results of management, as regards costs and the various activities.

Analytical accounting, also known as industrial or cost accounting, deals with analysing, processing and integrating data from general accounting with other information, in order to make them available for supporting decision-making processes, for the activities of planning, scheduling and management control.

The analytical accounting takes up the costs and revenues recognized from the general accounting, integrating the records and referring to the internal management: the information is reclassified by cost centres, which group all the financial movements with the same destination, i.e. the same user.

Cost centre is the elementary unit and can be a department or part of it, an operating group consisting of machines and men, any operating unit defined with the aim of attributing costs.

In the hospital setting, a cost analysis can prove to be very useful, especially for verifying the economic results achieved, i.e. the ratio between the costs incurred for the Diagnosis Related Groups (DRG) produced, thus assessing the degree of efficiency of the individual operating units and the structure as a whole.

A further consequence of the DRG-based mechanism is the focus of the information on the cost of the individual service, rather than on the costs produced by the single autonomous decision-making unit or by the single cost centre.

Only by attributing a value to the employed resources and to the activities performed in the context of a particular hospitalization, by identifying a cost capable of summarizing what has actually been spent, it will be possible to make a comparison with the obtained revenue, considering the DRG as the instrument that identifies the amount covering the incurred expenses. Therefore, it is more and more important being able to understand "how much is it?" to a hospital or an “Azienda Sanitaria Locale” (ASL) all the activities necessary for the care of a patient.

This economic and managerial information can be handled in an optimal way through tools designed by data scientists. In fact, BI tools can be fundamental for managing the large amount of produced data both economically and clinically, combining also these two aspects and thus providing managers with strategic information for corporate management.

2. The workflow of analysis and the integrated platforms

2.1. The workflow of BI

A BI system should meet functionality and design requirements that go far beyond those of a normal reporting environment, part of a management application [12-15]. Particularly, a BI system should have the following characteristics:

- *Ease of use*: the data must be simple to read and interpretable and it should be possible to easily navigate through them.
- *Velocity*: it is necessary to be able to process large amount of data, obtaining results in a very short time using modelling, storage and indexing techniques of data oriented to analysing rather than updating;
- *Integration*: integrating data coming from different sources, inside and outside the company. This process must be reliable and tested, so that users can rely on the data in the DW. If the data is not clean and reliable, it must go through a cleansing and certification process before being entered into the DW;
- *Data logging*: maintaining the history of changes which some attributes undergo in order to allow contextualized historical analysis;
- *Identifying trend and abnormalities*: it is necessary to make it easy to identify trends in the data by comparing different periods and products. These operations are possible only with the use of drill-down / roll-up (visualization of data at different levels of detail) and slice & dice (change of the analysis dimensions on the two axes);
- *Subject orientation*: the data should be shown to obtain the complete vision of a business process (supply chain, sales,

quality, ...), crossing the boundaries of the individual areas of the management systems;

- *Simulating scenarios*: in some cases (budgeting, forecasting and planning applications) it must be possible to set scenarios and then compare them with the real values ("actual");
- *Independence from the ICT workers:* the analysis and reporting tools must allow end users to create the reports they need by themselves;
- *Flexibility:* capacity to face and exploit the evolutions of operational systems and analysis' requirements within the company's reality
- *Security:* it must be possible to control the access to data, which includes in many cases highly confidential information, in a tight and flexible manner.

In figure 1 the so-called BI pyramid with its analysis steps is shown.

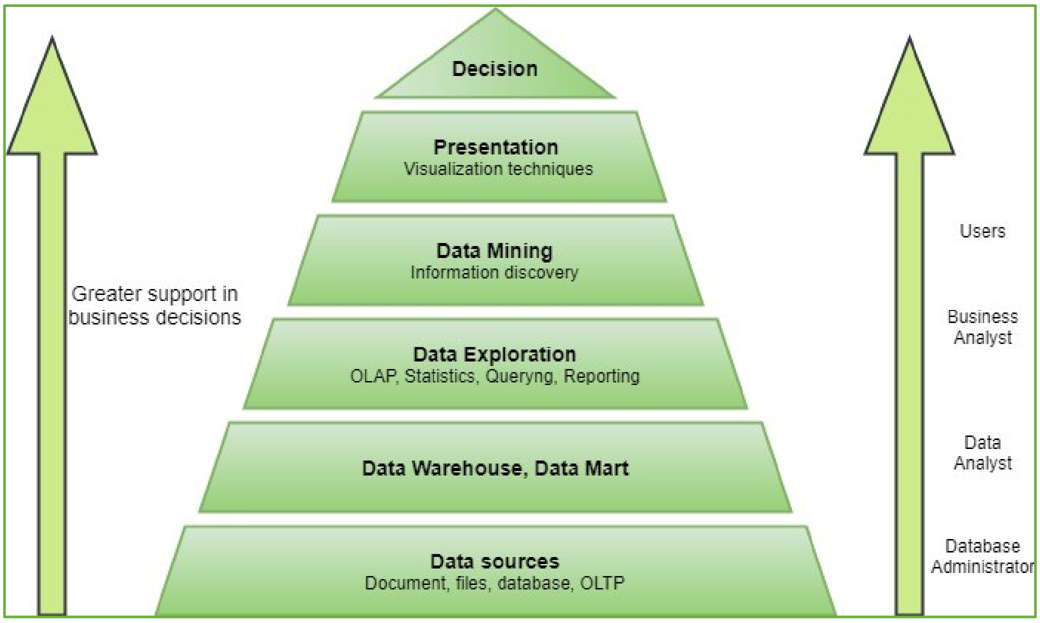


Figure 1. The workflow of BI, starting from data source to data analysis.

The DW is a software tool with the purpose of storing all data of business interest, whether they come from management systems or from external sources, in an optimized way for carrying out queries and not for writing. The information sources, in fact, are generally internal, coming from company information systems and integrated with all the others according to the needs. However, information from external sources such as customer base requests, shareholder estimates, technological / cultural trends or others may also be used. Each BI system has a specific goal that derives from the vision and objectives of a company's strategic management. For this reason, data structures alternative to those of operational databases have been conceived; while these are based on concepts and relational rules (Entity-Relationship), DWs are generally based on the dimensional model or Star-Schema, optimized to respond quickly to various types of queries.

The activity of Data Warehousing, meaning the building and the management of DW, involves several phases:

- Identifying the source data;
- Extraction, transformation and loading data (known as ETL workflow);
- Using a Database Management System (DBMS) to handle the DW;
- Employing BI tools to access the data.

Reporting systems are developed in complex areas that have provided the DW as the solution. One of the purposes of a DW process is to structure a hardware-software information context capable of responding to the needs of the organizational scenarios in the broadest sense.

As the stored data available to organizations grows, the advantages of centralized document processing are revealed in the execution times of the individual reporting documents: the particular hardware configuration of the workstations, on which the system resources are hosted physically, allows the optimization of the requests to the system and reduces the load, compared to the situation in which single users search for information on the system individually. A document, after being elaborated and generated, is validated by the appropriate structures and is distributed (and updated periodically) to the members of the organization who become its users.

The document produced is called report and is presented as a combination of tables and graphs that present the relevant measures for the various phenomena analysed, disaggregated and deconstructed according to needs. These measures form a common basis for subsequent analyses.

2.1.1. Power BI

Power BI is an open source software of BI that can be used both in desktop version and in a web-based and cloud version by paying a fee [16].

The tool can be connected to multiple data sources (from simple Excel to relational databases) and allows their aggregation in order to present and then display them in an intuitive way. This aggregation makes it possible to obtain useful information for the company management both at a clinical and administrative level and to find quantitative feedback with an information that could be only perceived in a qualitative way.

The connection with multiple data sources takes place thanks to the presence of specially developed connectors that make it possible to easily integrate the proposed tool. Furthermore, the data can be displayed in the most disparate ways (tables, histograms, linear charts, cakes, etc.) and with the definition of specific filters on fields chosen by the user.

Moreover, it is possible to create customizable dashboard templates (Figure 2) based on specific needs. The dashboards are dynamic and interactive, they update automatically when the user selects from the filters previously created in the template or specifically inserted in the dashboard according to the desires of the moment.

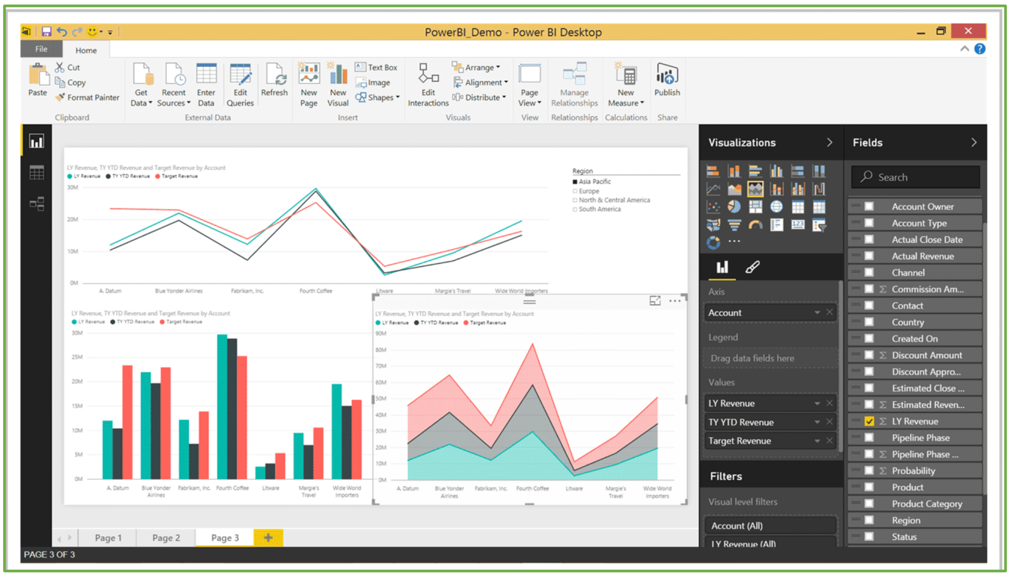


Figure 2. Example of interactive dashboard designed on Power BI.

Power BI allows, in an automated way, the geo-referencing of the data since it already has inside the ArcGIS maps and the connections with Google maps. They can be very useful for the analysis of georeferenced data in a given territory, mixed with interactive graphs.

Power BI is a data-driven platform made up of modules; it allows:

- The creation of new procedures for the management of particular processes, meaning that it is possible to add new procedures and then expand reporting and interfaces;
- The modification of an interface already created to make it customized for the user. For example, it is possible to modify a template to make it easier during the use in relation to the type of user;
- The creation of new interfaces that integrate with the old ones already existing.

Finally, the software satisfies the requirement of some known standards by applying:

- "ISO 13407: Human-centred design processes for interactive systems",

- “ISO TR 16982: Usability methods supporting human centred design”,
- “W3C Guidelines (Usability and Accessibility)”,
- “Nng – Nielsen Norman Group”.

That’s why the system has the following characteristics:

1. *Learnability*: a user is able to learn the functionality of the interface, easily carrying out the basic operations, even without having ever seen it before;
2. *Efficiency*: a user who knows the health process can carry out all the operations on the system in a simple and intuitive way;
3. *Memorability*: if a user previously uses the system, he will then find it easy to use it thanks to the features of the user-friendly interface;
4. *Errors management*: when making an operation that involves a transaction on the database, there are always alerts that warn about the type of operation that is performed. Furthermore, it is possible to remedy any errors made during the data entry phase in the database with the modification operations;
5. S*atisfaction*: using the system should be pleasant;
6. *Accessibility*: the system meets the needs of easy access based on user groups, the choice of input and output systems and the consistency in customization.

2.2. The workflow of machine learning

Before building a model and evaluating its performance, an understanding and preparation of the available data is essential. Therefore, scientists refer to a process known as Knowledge Discovery (KD) in Figure 3:

- The first step is the identification of the application domain and the goal to be achieved, as well as to fully define the problem to be addressed.
- The second step consists in preparing the data, and it involves:
    a. Data selection: the dataset is defined by selecting a subset of the most representative variables;
    b. Cleaning and pre-processing: they are operations aimed at eliminating noise caused by any anomalies recorded in the dataset, such as errors, missing values, inconsistent data, and any outliers;
    c. Transformation: it is a process through which the data undergoes normalization, dimensioning and selection operations, in order to obtain more practical manipulation.
- Now, it is possible to apply data mining techniques by identifying:
    a. The type of problem (classification, regression, clustering, etc);
    b. The most reliable algorithm for the problem;
    c. Parameters and hyperparameter setting which guarantee the best performance of the algorithms. This phase can be repeated until the best results is found.
- The fourth step consists in the assessment and interpretation of the results obtained by applying and implementing the algorithms. The evaluation of the models is made on the basis on specific criteria related to the examined problem; thus, it is fundamental the correct choice of the evaluation metrics.

At the end of this analysis, it will be possible to conduct a consolidation and use of the acquired knowledge.

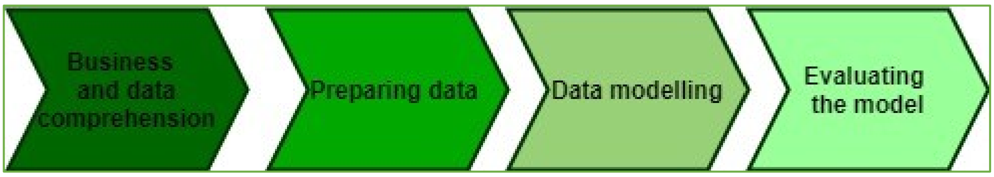


Figure 3. Workflow of the KD process.

2.2.1. The machine learning: validations and algorithms

Machine learning problems can be divided into [17]:

- Supervised learning: the data supplied to the model present the class to which they belong, in a column called target. The dataset will be made up of N observations, described by m features and a target column (label). In turn, supervised learning can be divided into two problems:
  - Classification: the data are represented by association with a classification label, or more simply a class. Starting from a set of observations whose class they belong to is already known, a model capable of making predictions is generated;
  - Regression: it is a statistical process that tries to establish the relationship between two or more variables, thus allowing prediction on data with non-categorical targets. Therefore, unlike the classification, which limits itself to discriminating data based on the classes and with a discrete value, the regression returns a continuous output starting from the input provided;
- Unsupervised learning: the data provided to the model do not have the target column. Moreover, in this case, unsupervised learning can be divided into problems:
  - Clustering: It aims at the selection and grouping of homogeneous records in a set of data, by means of measures relating to the similarity between elements. In many approaches, this similarity is

    conceived in terms of multidimensional distance and the goodness of the groupings depends on the chosen metric;
    - Association Rules: it is a method that aims to extract information on hidden relationships between the variables. It represents an effective method for analysing data without a priori knowledge of the correlation between them.

Focusing on the machine learning analysis of supervised learning, it is carried out on a set of samples, which are divided into training sets and test sets: the model is built and implemented using the first, while the performance of the implemented model is assessed through the latter.

There are 3 types of cross-validation (CV) that can be adopted for both classification and regression problems [18]:

- Hold-Out: it consists in dividing the dataset into two sets, the first for training, used for the adaptation and training of the model, and the other one for testing, used to evaluate the model.
- K-Fold CV: in this approach the set of observations (the dataset) are randomly divided into k data groups, or folds; the training and test procedures of the model are repeated k times using k-1 fold for training and one fold for the test. From a computational point of view, this technique turns out to be more expensive but more reliable than hold-out, if the dataset is not big enough.
- Leave-one-out CV: it is similar to k-fold but it divides the dataset into as many folds as the number of records in the dataset. Subsequently, only one element is removed from the dataset at a time and used to test the model while all the other data constitute the training set to train the model. One of the advantages of this particular technique is the non-overestimation of the error. This technique is more computationally expensive than both hold-out and k-fold but it is useful for small datasets.

Moreover, data pre-processing techniques are frequently used to identify which are the most important features that can maximize the accuracy of the classifier in its task; the reference is to the Wrapper [19].

Finally, the main algorithms used in the literature validations are listed, briefly described and will be mentioned in the following paragraphs:

- Naïve Bayes (NB) is a supervised learning algorithm suitable for solving both binary (two-class) and multi-class classification problems. The main peculiarity of the algorithm, in addition to making use of the Bayes theorem, is certainly being based on the hypothesis of non-correlation of the characteristics, that is, the classifier assumes the independence of the predictors within each class [20].
- The Decision Tree (DT) is a very simple and effective algorithm. Within each DT, an internal node is associated with a particular "question" about a feature. From this node, as many arcs branch off as there are possible values that the characteristic can assume, until the leaves that indicate the category associated with the decision are reached. The division criteria from each node to get to the leaves can be different, although they do not particularly influence the results [21].
- The K nearest neighbour (KNN) is an instance-based algorithm, based on distances, it is not necessary to make assumptions about the underlying distribution of the data, usually the similarity is calculated through the Euclidean distance. Distances are computed in the training data and then applied in the test, the shorter the distance and the greater the similarity between the observations. The algorithm provides for setting a parameter k which identifies the number of closest neighbours to be considered in order to assign the class to the new element [22].

- The Support Vector Machine (SVM) obtains maximum effectiveness in binary classification problems and is based on the idea of finding the best hyperplane that divides a data set into two classes. The classifier starts looking for a hyperplane that linearly separates the two classes and, if more than one existed, it would find the one that has the highest margin with the support vectors to improve the accuracy of the model. If such a hyperplane did not exist, SVM would use a non-linear mapping to transform the training data into a hyperspace where the hyperplane can, however, be identified [23].
- In order to obtaine better results, on the so-called “weak learners”, ensemble learning techniques have been implemented:
    - Bagging: Bootstrap AGGregatING tries to reduce overfitting problems by training different classifiers on subsets of the training set and finally performing a majority vote; when applied to the DT, it generates a powerful algorithm known as Random Forests (RF) [24].
    - Boosting: it tries to reduce overfitting problems by training different classifiers on non-random sub-parts of the training set; it chooses, in fact, the records that have not been correctly classified. Famous variants are those applied to DT (ADA-B) or together with Bagging as Gradient Boosted Tree (GB) [25-26].

There are also many form of neural networks, they are structures simulating the one of the human brain: they are divided into layers of neurons. Multilayer Perceptron (MLP) is the most employed in this work.

2.2.2. Knime Analytics Platform

Knime Analytics Platform (figure 4) is an open source platform for data analysis, integration and reporting, with modules on machine learning and data mining. It allows the creation of workflows by using an intuitive graphical interface and drag and drop, choosing from more than 2000 nodes that allow the modelling of each phase of the analysis. Built by a team of engineers led by Michael Berthold initially for pharmaceutical industries [27], it was born with the intention of providing a modular and highly scalable platform to be used in any application area, which is why it has had great diffusion in various areas such as customer data analysis, BI, text mining and financial data analysis .

Figure 4. Knime Analytics Platform logo.

Knime is written in Java and based on Eclipse, which makes use of the extension mechanisms to add plugins necessary for the realization of additive functionalities.

The characteristics of Knime are:

- Using data coming from any sources. It is possible to:
    - Import and combine text formats (CSV, PDF, XLS, JSON, XML, etc) and unstructured data (images, documents, networks);
    - Integrating data from Oracle, Microsoft SQL, Apache Hive and others, through the connection to external sources and hosts;

    - Accessing and obtaining data from sources such as Twitter, Azure, Google Sheets.
- Data modelling. It allows to:
    - Calculate statistics, including mean, quantile, standard deviation, or apply statistical tests to validate a hypothesis
    - Converting data through reordering, filtering, joining even on huge amounts of data;
    - Cleaning data through operations of normalization and missing value management;
    - Extract and select the features, in order to prepare and optimize the dataset for subsequent use in machine learning.
- Supporting machine learning and artificial intelligence. It allows to:
    - Build machine learning models for classification, regression, clustering, dimension reduction, by using advanced algorithms, including deep learning ones;
    - Optimize the performance of models through boosting, bagging, hyperparameters setting and ensemble learning;
    - Validating and evaluating models through metrics such as accuracy, Area Under the Curve Receiver Operating Characteristics (AUCROC), sensibility, specificity, recall, precision and CV, feature filtering, features selection, etc;
- Focusing and sharing results. It is possible to:

    - Visualize the data through simple graphs, such as bar charts and scatter plots, or advanced, with the possibility of customizing them according to the needs;
    - Visualize statistical results through tables and summaries;
    - Save, store and export the data and the results of the analysis in several formats and database.

KNIME, as an open-source software, focuses on integration and allows third-party developers to easily integrate their tools and make them interoperable, regardless of their domain. In addition, it also provides users with numerous tools that, despite not being strictly designed for the scientific and biological field, can be important for data acquisition, visualization and statistical analysis. Several databases allow the user to upload their own data from heterogeneous sources and script support allows developers to program nodes for data processing in their preferred language (Python, R, etc). Moreover, WEKA, Python, Java, MATLAB and many implementations of the most popular algorithms and data mining tools, which enrich the toolbox for analysis, are available as plugin [28].

KNIME operates with a table-by-table process, the advantages are:

- Performing multiple iterations on the same data, which is fundamental for data mining algorithms;
- Visualizing intermediate results between node connections even before the workflow is executed;
- Restoring and restarting the workflow from any node [29].

An example of workflow on Knime Analytics Platform is shown in figure 5.

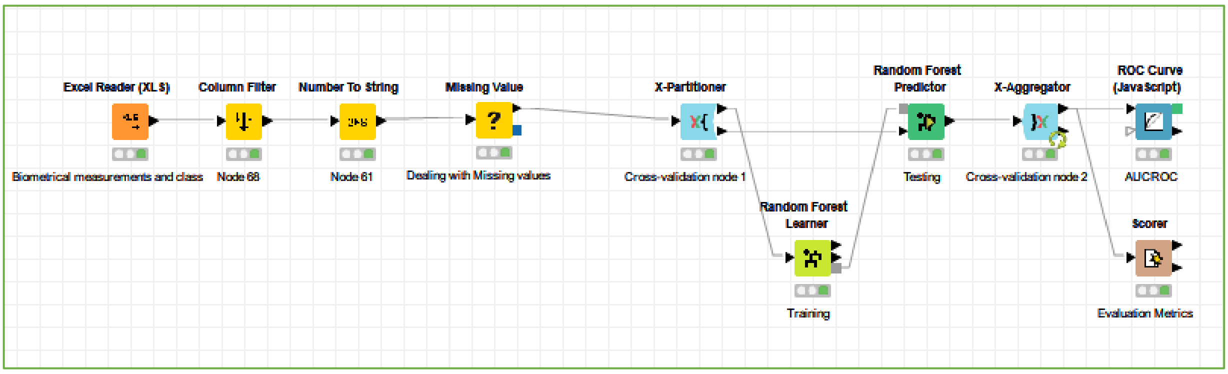


Figure 5. Example of workflow for an analysis on Knime Analytics Platform.

3. Validating tools

Both tools have been extensively validated, Power BI with a regional project and with an application to the current pandemic context and Knime analytics platform through scientific publications in international literature in several fields of medicine. The results achieved with these tools are shown in the following paragraphs.

3.1. Co.S.A. project

The project called "Correlazione Salute Ambiente" (Co.S.A.) had the goal of creating a structured BI platform to allow the collection of Big Data from heterogeneous data sources substantially linked to health, environmental and socio-economic information.

Figure 6 shows the project workflow.

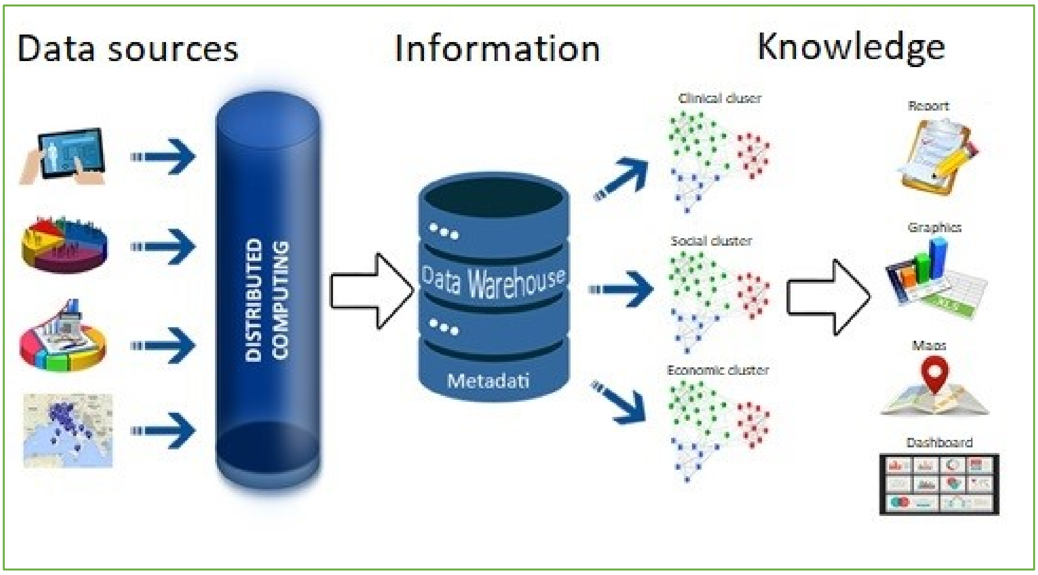


Figure 6. The main parts and the aims of Co.S.A. project.

The final result was achieved through the use of Power BI, it allowed to obtain a quick manipulation of the data collected in the database, having an organized set of information elements ready to be included in the modules dedicated to the management of the respective reference datasets. Furthermore, it was possible to manage all the manipulation and aggregation operations of the data and, consequently, their representation on maps and graphs.

The system allowed the visualization of the data stored in the DW by providing special views on the graphic interfaces. In particular, the end user had the opportunity to:

- Visualizing environmental data with georeferenced maps.
- Understand the places where biological samples were taken thanks to the maps showing the sampling points and the levels of heavy metal pollution that were performed;
- Visualizing health data through scale maps with the density of the population per each city with the number of people affected by chronic respiratory disease.

3.2. Machine learning in Cardiology

Cardiac diseases are among those with the highest mortality in the world and there are datasets with many variables that need to be evaluated for the management of patients with these pathologies. It was, therefore, considered useful a validation of the platform in this context.

Several applications were conducted:

1. The aims of the present study were to test data mining and machine learning tools and to compare different algorithms of supervised learning in a large cohort of Italian patients with suspected or known coronary artery disease (CAD) who underwent stress SPECT myocardial perfusion imaging (MPI). The accuracy of the results was greater than 95% showing that data mining was feasible within the daily process for the institutions in which clinical variables are systematically registered along with simple diagnostics [30].
2. Two biomedical technologies (Cadmium–zinc–telluride (CZT) and Anger camera) were compared for their ability to diagnose CAD. The CZT camera showed improved diagnostic accuracy and increased sensitivity in two studies:
   - 2.1. With female patients [31], whose breast could lead to reduce soft tissue attenuation;
   - 2.2. With larger and more generalized cohort of patients [32].
3. In the present study, I compared the integration of the 11 Nonlinear Trimodal Regression Analysis (NTRA) parameters to classify elderly at risk for cardiac pathologies using multivariate logistic regression modelling and three different tree-based machine learning algorithms. Starting from mid-thigh CT images, radiodensitometric profiles of muscle, fat and connective tissues were used to stratify the risk of patients with cardiological and cardiovascular diseases, achieving accuracies of over 95%. [33].

4. A final more methodological application aimed at testing the effect of linear discriminant analysis alone and combined with principal component analysis for the detection of coronary pathologies. The accuracy of the models did not have strong differences, settling respectively at 84.5% and 86% [34].

The first three papers started from a dataset of about 10,000 patients who were screened for the presence or suspected presence of CAD, in particular they were all subjected to SPECT MPI. Therefore, the first study represented SPECT MPI-based investigation on the largest dataset in literature. Table 1 shows the results achieved in the classification.

Table 1. Overall scores (accuracy=acc., sensitivity= sens., specificity=spec., recall=rec., precision=prec.) in percentage (%) for the classification of CAD.

Algorithms	Acc.	Error	Prec.	Rec.	Spec.	Sens.
RF	96.9	3.1	94.8	97.0	96.7	97.0
GB	96.8	3.2	94.9	96.7	96.9	96.7
DT	85.9	14.1	73.1	99.3	77.8	99.3
KNN	87.5	12.5	87.7	79.2	90.1	79.2
NB	83.0	17.0	70.9	89.0	77.9	89.0

In the study at point 3, the NTRA parameters, based on the radiodensitometric distributions of absorption of the rays in CT, had previously been used to study sarcopenia and quantify the following muscle degeneration; they were used in this study to predict the possibility of contracting heart, coronary and cardiovascular diseases in elderly subjects.

AGES dataset was used [35], it was developed in to study genetic pathologies in the Icelandic population. The main results and the workflow of the study are shown in figure 7 and summarized in table 2.

Table 2. Overall scores (accuracy=acc., sensitivity= sens., specificity=spec., recall=rec., precision=prec.) in percentage (%) for classifying coronary heart disease (CHD), cardiovascular disease (CVD) and chronic heart failure (CHF).

	Algorithms	Acc. Mean	Acc. Max	Sens.	Spec.	Rec.	Prec.	AUCROC
CHD	GB	75.9	77.7	70.0	81.7	70.0	79.3	0.864
	RF	85.0	87.4	81.7	88.4	81.7	87.6	0.936
	ADA-B	79.5	82.2	74.9	84.1	74.9	82.4	0.873
CVD	GB	73.1	75.7	67.1	79.1	67.1	76.2	0.834
	RF	82.1	83.9	78.8	85.5	78.8	84.5	0.914
	ADA-B	70.2	77.0	63.3	77.2	63.3	73.5	0.766
CHF	GB	88.6	90.3	85.0	92.1	85.0	91.5	0.962
	RF	95.9	96.5	95.0	96.9	95.0	96.8	0.994
	ADA-B	94.0	95.4	92.1	95.8	92.1	95.7	0.987

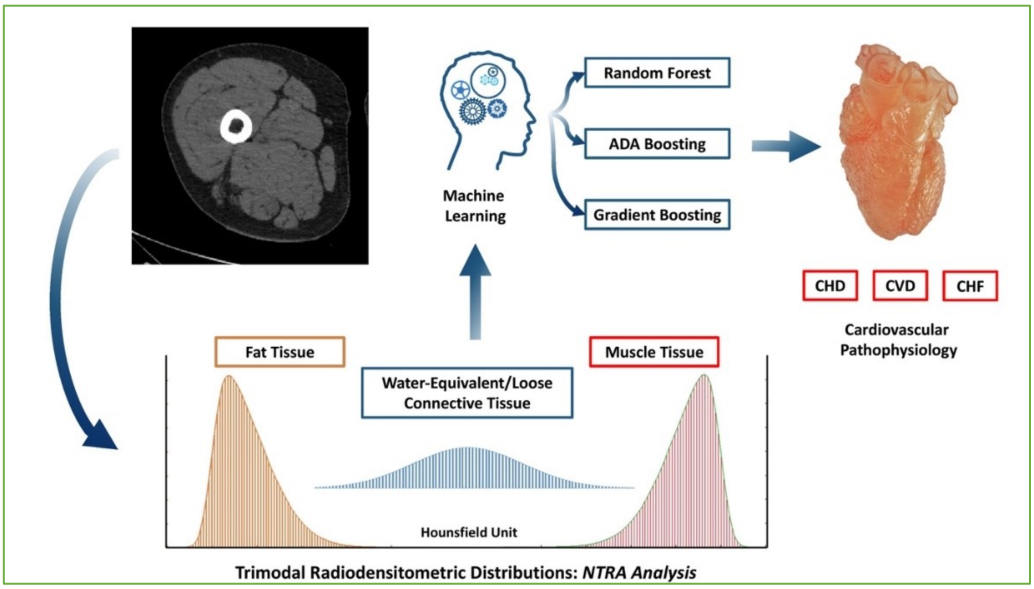

Machine Learning
Random Forest
ADA Boosting
Gradient Boosting
CHD
CVD
CHF
Cardiovascular Pathophysiology
Fat Tissue
Water-Equivalent/Loose Connective Tissue
Muscle Tissue
Hounsfield Unit
Trimodal Radiodensitometric Distributions: *NTRA Analysis*

Figure 7. The workflow of the machine learning study in cardiology, based on the NTRA parameters for evaluating the risks associated to cardiac and cardiovascular diseases.

The results of the study number 4 are represented in table 3. In cardiology, the variables to be considered for each patient are several; so, it was decided to evaluate the impact of a feature projection algorithm to reduce the number of variables to be given as input to machine learning algorithms.

Table 3. Overall scores (accuracy=acc., sensitivity= sens., specificity=spec., recall=rec., precision=prec.) in percentage (%) for classifying CAD with only a linear discriminant analysis (LDA) and then combined with a principal component analysis (PCA).

Algorithms	Acc.	Error	Rec.	Prec.	Sens.	Spec.
LDA	84.5	15.5	62.8	94.2	62.8	97.7
LDA and PCA	86.0	14.0	65.4	96.2	65.4	98.4

3.3. Machine learning in Neurology

In the neurological field, Parkinson's disease (PD) is a neurodegenerative disease, with a slow but progressive evolution, which mainly involves some functions such as the control of movements and balance. The disease is part of a group of pathologies called "Movement Disorders" and among them it is the most frequent. PD clinical picture is characterized by a combination of motor (bradykinesia, resting tremor, rigidity, gait and stability impairment) and non-motor symptoms (cognitive impairment, depression, psychosis, constipation, etc.) that worsen as the disease advances.

The aim of the research was to use the spatial and temporal parameters acquired through gait analysis, a three-dimensional, computerized and non-invasive examination of the walk, to make differential diagnosis of parkinsonism and to correlate gait patterns to PD’s symptoms.

It allows to monitor patient's movement and to quantitatively measure several aspects of gait that becomes fundamental in the evaluation of his functional limitation. Through a multifactorial analysis, it is possible to define, by using sophisticated equipment, integrated with each other, the ambulatory pattern of the subject in question. Posture and movement, in fact, are the result of the interaction of three main physiological systems: the nervous system, the musculoskeletal system and the sensory system. The evaluation of the characteristics of posture and movement, as well as their variations with respect to a normal situation, can be of enormous utility in the clinical field for the diagnosis of particular pathologies affecting one of the systems involved, as well as for the planning and control of specific rehabilitation treatments. The BTS Gait Lab was used to carry out the acquisitions of all patients; the laboratory is composed of 6 infrared cameras, 2 video cameras, 22 reflective markers (coated with silver powder), two form platforms, 4 pairs of surfaces electromyographs and a computer.

The Davis protocol was chosen [36], it involves 4 phases:

1. Anthropometric measurements of patients;
2. Positioning of 22 reflective markers on patients;
3. Standing phase;
4. Walking phase.

Each acquisition was conducted in three conditions:

- Normal gait (GAIT),
- Walking by carrying a tray with two glasses full of water, corresponding to a dual motor task (MOT);
- Walking while serial subtracting 7s, corresponding to a cognitive dual task (COG).

After having acquired the patients and processed the reports (the processing interface is shown in Figure 8), the spatial and temporal parameters of the walk were obtained (shown in table 4) and various purposes were pursued.

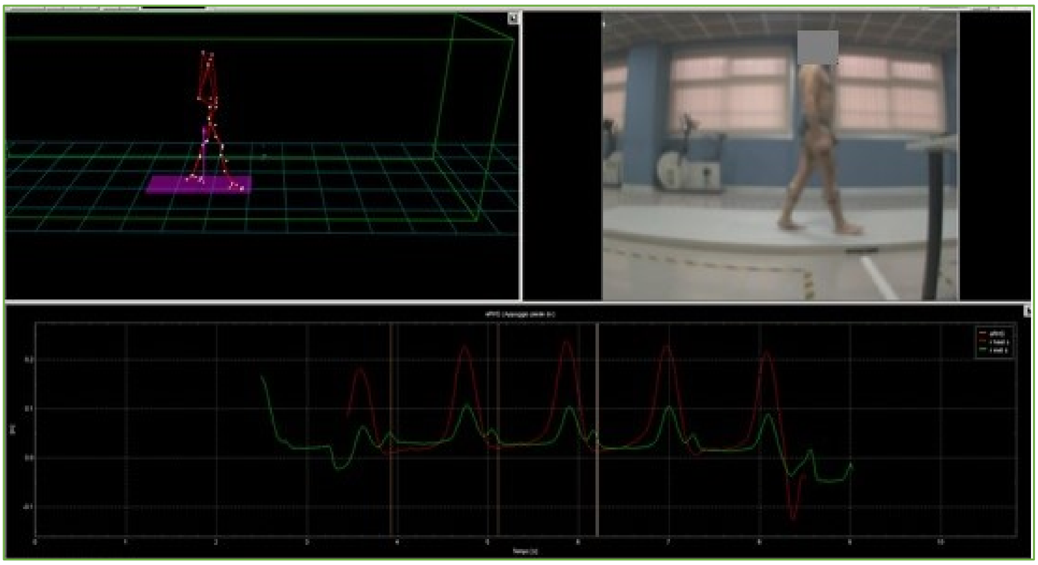


Figure 8. Elaboration of a signal with the three-dimensional reconstruction of the patients and the video recording provided by BTS system.

Table 4. Spatial and temporal parameter of gait.

Cycle duration (s)	Mean velocity (m/s)
Stance duration (s)	Mean velocity (%height/s)
Swing duration (s)	Cadence (step/min)
Variability of swing's duration (s)	Cycle length (m)
Stance phase (%)	Cycle length (%height)
Swing phase (%)	Step length (m)
Single support phase (%)	Variability of the length of step (m)
Double support phase (%)	Step width (m)

The first was to distinguish patients with PD for several years from patients with Progressive Supranuclear Palsy (PSP), through RF and GB an accuracy of 86% and a sensitivity greater than 90% for atypical parkinsonism was obtained [37]. Table 5 shows the main results achieved.

Table 5. Overall scores (accuracy=acc., sensitivity= sens., specificity=spec., recall=rec., precision=prec.) in percentage (%) for the classification of PSP and PD.

Groups	Algorithms	Acc.	Error	Rec.	Prec.	Sens.	Spec.
PSP	**RF**	86.4	13.6	92.6	92.6	92.6	96.3
	GB	84.0	16.0	96.3	86.7	96.3	92.6
De Novo PD	**RF**	86.4	13.6	96.3	81.3	96.2	88.9
	GB	84.0	16.0	88.9	80.0	88.9	88.9
Stable PD	**RF**	86.4	13.6	70.4	86.4	70.4	94.4
	GB	84.0	16.0	66.7	85.7	66.7	94.4

Diagnosing the form of parkinsonism is not an easy task for neurologists, usually some attempts are made with drug therapies and evaluating patient's response to the treatment. The proposed approach, combining the use of gait analysis and machine learning, could allow doctors to have a clinical DSS.

In the same context, the platform was then used to investigate how walking, through the spatial and temporal parameters of gait, can be used to detect even hidden symptoms of PD (for example, mild cognitive impairment (MCI), the results are in table 6, and the freezing of gait, the results are in table 7) [38-39].

Table 6. Overall scores (accuracy=acc., sensitivity= sens., specificity=spec., recall=rec., precision=prec.) in percentage (%) for detecting MCI.

Alg.	Task	Acc.	Sens.	Spec.	AUCROC
DT	**GAIT**	75.0	73.5	76.5	0.789
	MOT	86.8	88.2	85.3	0.841
	COG	77.9	70.6	85.3	0.771
RF	**GAIT**	76.5	82.4	70.6	0.858
	MOT	75.0	70.6	79.4	0.858
	COG	82.4	85.3	79.4	0.885
KNN	**GAIT**	83.8	88.2	79.4	0.900
	MOT	60.3	50.0	70.6	0.761
	COG	79.4	73.5	85.3	0.818

Table 7. Overall scores (accuracy=acc., sensitivity= sens., specificity=spec., recall=rec., precision=prec.) in percentage (%) for detecting anamnestic freezing of gait.

GB				
Task	**Acc.**	**Spec.**	**Sens.**	**AUCROC**
GAIT	0,87	0,77	0,97	0,97
MOT	0,87	0,81	0,94	0,98
COG	0,82	0,81	0,84	0,87
TOT	0,94	0,94	0,94	0,94
ADA-B				
Task	**Acc.**	**Spec.**	**Sens.**	**AUCROC**
GAIT	0,81	0,75	0,88	0,90
MOT	0,88	0,75	1,00	0,92
COG	0,88	0,75	1,00	0,80
TOT	0.88	0,88	0,88	0,88

The former study showed that a machine learning method using gait analysis features displays good accuracy, specificity and sensitivity in detecting the presence of MCI in Parkinsonian patients, thus supporting the existence of specific connections between gait and cognition in PD.

As regards the latter study, first, the employment of algorithms and analytics platforms can help neurologists in clinical decision-making since some patients may suffer from freezing of gait, despite not showing the phenomenon during medical visits: a machine learning analysis combined with a gait exam can detect these mild freezers. Furthermore, considering the consequences of this problem in the life of elderly patients (falls and fractures), this application would help to identify patients at increased risk of falling, thus reducing morbidity and mortality.

3.4. Machine learning in the radiomic process

The platform was also used for radiomics studies. Indeed, machine learning can also be applied to medical images through the quantitative analysis of radiomics features extracted through the so-called texture analysis.

Here are some of the studies conducted in this context:

- In a first study, the aim was to evaluate whether a machine learning analysis, employing magnetic resonance imaging-derived texture analysis features, could be useful in assessing the presence of placental adhesion disorder in patients with placenta previa. The hypothesis was that texture analysis features may reflect histological abnormalities underlying placental adhesion disorder in patients with placenta previa thus helping in differentiating positive from negative cases. Among the tested algorithms, the k-NN obtained the highest accuracy (98.1%), precision (98.7%), sensitivity (97.5%) and specificity (98.7%) values, while exploiting the lowest number of features (26). [40]. The results are shown in table 8.

Table 8. Overall scores (accuracy=acc., sensitivity= sens., specificity=spec.) in percentage (%) for detecting patients with placenta previa.

Algorithms	Acc.	Prec.	Sens.	Spec.
RF	95.6	97.4	93.7	93.7
KNN	98.1	98.7	97.5	98.7
NB	80.5	77.3	86.1	75.0
MLP	88.6	84.9	92.4	83.8

- In a different study, a machine learning analysis was performed using Computer Tomography-derived texture analysis features in patients with histologically proven cancers of the oral cavity and oropharynx squamous cell carcinoma, aiming to assess whether this approach may predict tumour grade and nodal status; in this case, regarding the classification of tumour grade, NB achieved the best accuracy (92.9%) and the highest AUROC (over 0.900). Concerning the classification of nodal status, NB overcame the accuracy of 90.0% [41]. Results are shown in tables 9 and 10.

Table 9. Overall scores (accuracy=acc., sensitivity= sens., specificity=spec.) in percentage (%) for classifying tumour grade in patients with cancers of the oral cavity and oropharynx squamous cell carcinoma.

	Algorithms	Acc.	Sens.	Spec.	AUCROC
No ensemble	J48	78.6	85.7	71.4	0.888
	MLP	78.6	85.7	71.4	0.816
	NB	92.9	85.7	100	0.980
	KNN	78.6	71.4	85.7	0.878
Bagging	J48	78.6	85.7	71.4	0.939
	MLP	71.4	85.7	57.1	0.852
	NB	92.9	85.7	100	0.857
	KNN	92.9	100	85.7	0.939
ADA-B	J48	85.7	85.7	85.7	0.939
	MLP	85.7	71.4	100	0.878
	NB	78.6	85.7	71.4	0.827
	KNN	85.7	85.7	85.7	0.867

Table 10. Overall scores (accuracy=acc., sensitivity= sens., specificity=spec.) in percentage (%) for classifying nodal status in patients with cancers of the oral cavity and oropharynx squamous cell carcinoma.

	Algorithms	Acc.	Sens.	Spec.	AUCROC
No ensemble	J48	90.9	80.0	100	0.900
	MLP	72.7	60.0	83.3	0.800
	NB	90.9	100	83.3	0.867
	KNN	81.8	100	66.7	0.900
Bagging	J48	81.8	80.0	83.3	0.917
	MLP	72.7	100	50.0	0.833
	NB	92.9	85.7	100	0.857
	KNN	72.7	100	50.0	0.900
ADA-B	J48	90.9	80.0	100	0.900
	MLP	72.7	80.0	66.7	0.650
	NB	81.8	60.0	100	0.900
	KNN	72.7	60.0	83.3	0.717

- Finally, the aim of the present study was to evaluate the feasibility of a combined texture analysis and machine learning approach on magnetic resonance imaging for the prediction of Fuhrman grade in renal cell carcinoma with the clear cell subtype. All the ensemble methods achieved an accuracy greater than 90%: ADA-B obtained the best accuracy (92.7%), recall and sensitivity (91.7%); Bagging showed the highest AUCROC [42]. Table 11 reports results on both training and test sets.

Table 11. Overall scores (accuracy=acc., sensitivity= sens., specificity=spec.) in percentage (%) for distinguishing tumour grade in renal cell carcinoma.

	Test on Real data		Test on Artificial data					
	Acc.	AUCROC	Acc.	Rec.	Prec.	Sens.	Spec.	AUCROC
J48	87.5	0.900	74.0	55.5	69.0	55.5	85.0	0.766
Bagging	62.5	0.733	90.6	88.9	86.5	88.9	91.7	0.952
RF	75.0	0.633	91.7	88.9	88.9	88.9	93.3	0.918
ADA-B	75.0	0.733	92.7	91.7	89.2	91.7	93.3	0.933

In figure 9 the workflow for the first and the third studies is shown, in the second one computer tomography images replaced magnetic resonance imaging while the other steps remained the same.

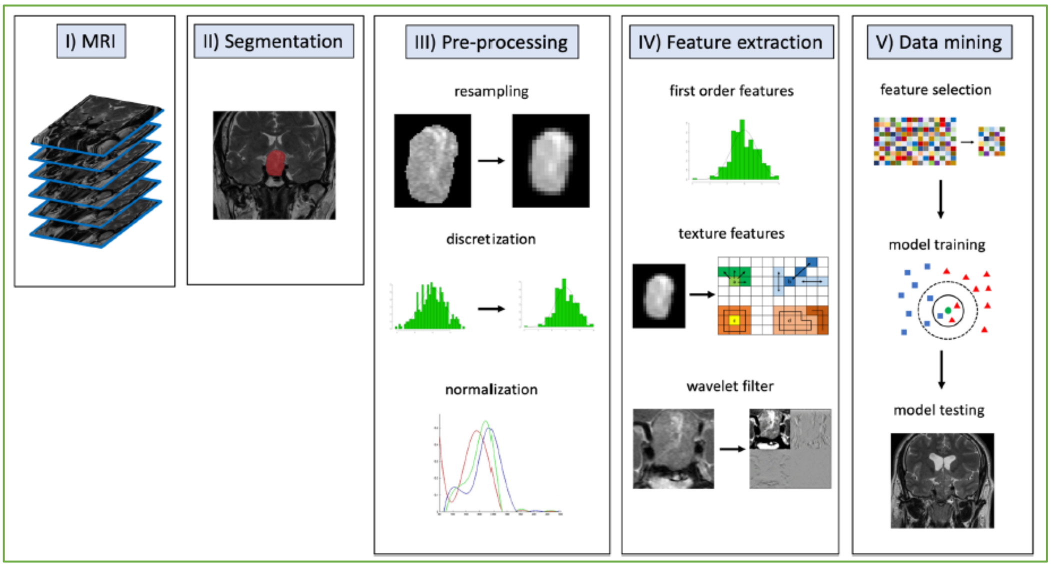


Figure 9. All the steps for a radiomic process.

3.5. Machine learning on biomedical signals

A part of the large amount of data generated in the healthcare sector certainly consists of biomedical signals: electrocardiogram, electroencephalogram, electro-oculography, cardiotocography, etc.

Two studies were conducted in this context:

1. The first study aimed at classifying the type of delivery of women (vaginal or caesarean section) starting from features extracted from a cardiotocographic signal [43].
2. The aim of the second study was to investigate the relationships between physiological values of congenital nystagmus affected people and features extracted from their electro-oculography through several machine learning algorithms and compute some evaluation metrics to compare the new results with the past ones without this approach. [44].

Fetal well-being and cardiotocography (figure 10) have been the subject of research since the middle of the last century; in fact, since then researchers have been looking for the best way to objectively interpret this signal and the features that can be extracted from it. The results in the distinction between Caesarean and natural birth through 17 features (in table 12 with their international wording) are reported in table 13.

Table 12. Features extracted from cardiotocographic signal.

FHR= fetal heart rate.

FHR mean	Variability index from RR series (index estimated by means of symbolic dynamics analysis)
Number of accelerations	Variability index from FHR series
Number of decelerations	Sample entropy (for quantifying the randomness or regularity/irregularity of the fluctuations of heart beats)
Number of Uterine Contractions	Minor axis of the data map obtained with the Poincaré technique
Presence of tachycardia	Major axis of the data map obtained with the Poincaré technique
Power in low frequency of FHR power spectrum (0.05 - 0.2 Hz)	Fractal dimension (estimated by means of Higouchi algorithm)
Power in high frequency of FHR power spectrum (0.2 - 1 Hz)	Number of pregnancy weeks
Sympatho Vagal Balance	Antepartum/Labour/In presence of uterine contractions
Short Term Variations	

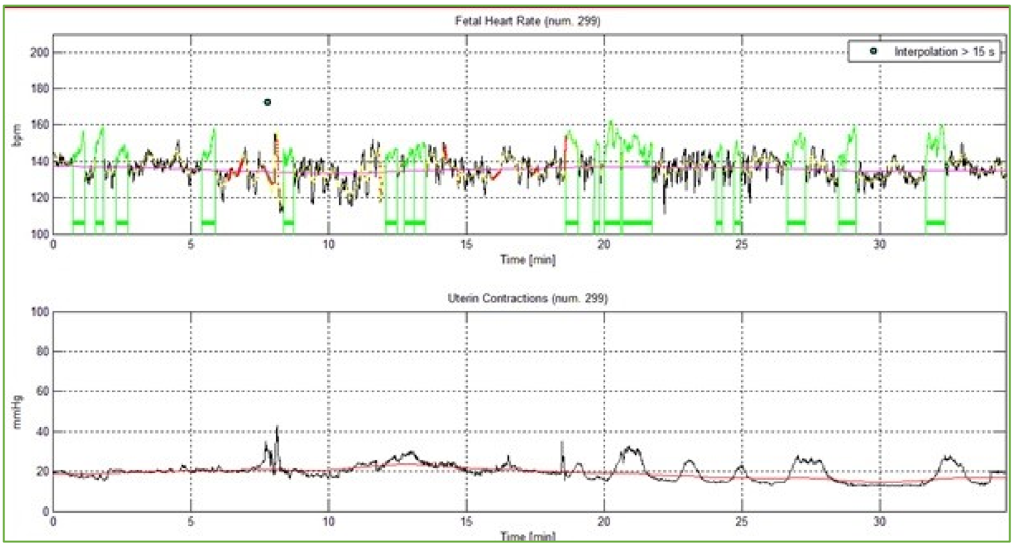


Figure 10. Cardiotocographic signal with Fetal Heart Rate and uterine contractions.

Table 13. Overall scores (accuracy=acc., sensitivity= sens., specificity=spec.) in percentage (%) for distinguishing natural delivery and caesarean section.

Algorithms	Acc.	Prec.	Sens.	Spec.	AUCROC
DT	82.5	84.3	80.1	85.0	83.9
ADA-B	89.6	92.6	86.1	93.1	95.5
RF	91.1	92.1	90.0	92.2	96.7
GB	87.5	87.1	88.1	87.0	94.9
Decorate	88.5	90.4	86.1	90.9	95.5

The purpose of the second study was to study through the regression the relationships between physiological values (visual acuity and variability of eye-positioning of patients) with congenital nystagmus and parameters (amplitude and frequency of nystagmus, period of foveation, amplitude and frequency of baseline oscillation) extracted from electro-oculography (figure 11); the results are reported in table 14.

Table 14. Overall scores for the regression of visual acuity (VA) and variability of eye-positioning (SDp).

Regression on VA						
	RF	**Logistic based on DT**	**KNN**	**GB**	**MLP**	**SVM**
R^2	0.85	0.72	0.67	0.82	0.83	0.68
Mean absolute error	0.05	0.06	0.06	0.04	0.10	0.48
Mean squared error	0.00	0.01	0.01	0.00	0.02	0.34
Root mean squared deviation	0.06	0.08	0.09	0.07	0.13	0.58
Mean signed difference	0.00	0.01	0.01	0.00	0.00	0.05

Regression on SDp						
	RF	**Logistic based on DT**	**KNN**	**GB**	**MLP**	**SVM**
R^2	0.65	0.62	0.74	0.68	0.72	0.79
Mean absolute error	0.32	0.34	0.27	0.31	0.10	0.09
Mean squared error	0.16	0.17	0.12	0.14	0.02	0.01
Root mean squared deviation	0.4	0.42	0.34	0.38	0.13	0.11
Mean signed difference	0.00	0.00	-0.04	-0.01	0.00	0.00

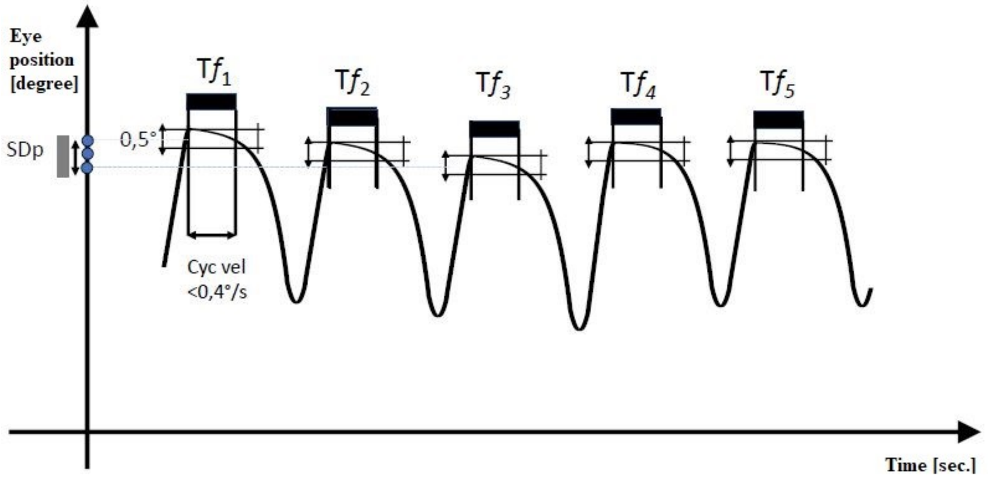


Figure 11. Electro-oculographic signal with the description of some features (variability of eye-positioning (SDp) and foveation period (Tf)).

3.6. Application on the actual pandemic context: COVID-19

3.6.1. The context

At the end of 2019, a new highly contagious virus completely unknown to our immune system started a circus in a remote region of the globe (Wuhan, China). In a few months, this virus, known as SARS-CoV-2 (below it will be referred to as COVID-19), has radically changed the global scenario from a social, economic and health point of view. The virus spreads so much in early 2020 that the World Health Organization has decided to officially declare "the public health emergency of international concern".

COVID-19 is a severe acute respiratory syndrome and is genetically closely related to the SARS-CoV-1 virus that causes SARS, which emerged in late 2002 in China. To date, various tools are used to diagnose this virus, oropharyngeal swabs and serological tests; over time, rapid tests were also developed to speed up population screening and allow tests to be performed in private centres. In fact, the swabs have been made available to private bodies for the diagnosis of the disease to the population only at the end of 2020.

According to the "art. 8 del Decreto-Legge 9 Marzo 2020 n.14" with the "Disposizioni urgenti per il potenziamento del Servizio Sanitario Nazionale in relazione all'emergenza COVID-19" and the document for the homogeneous application of "Decreto-Legge 9 marzo 2020" released by the Health Committee, approved on the 16th of March 2020, the special units "Unità Speciali di Continuità Assistenziale" (U.S.C.A.) must be set up in the health districts "Aziende Sanitarie Locali" (ASL) in order to implement the management of the health emergency for the COVID-19 epidemic and allow the General Practitioner "Medico di Medicina Generale" (MMG) and the Free Choice Paediatrician "Pediatra di Libera Scelta" (PLS) to guarantee ordinary care activities. The purpose of the U.S.C.A. is to ensure home care for COVID-19 patients who do not need hospitalization, including the administration of drugs at home.

To satisfy the health and social needs of the population, the identification of the needs and the objectives to be achieved by ASL Napoli 1 was organized in 10 districts (from 24 to 33). In the changed emergency health context, the U.S.C.A. act as an inter-district body that is inserted as an operational support in the management of the emergency. In this context, in Gesan S.r.l. (San Nicola la Strada, Caserta, Italy) a computerized platform, Integrated Covid-19 Surveillance (SIC) has been developed. It allows to handle the management of the patient with suspected positive COVID-19 aiming to guarantee the best prognosis, the best clinical course and to reduce the hospitalization rate.

To enhance the capabilities of the SIC, the use of three platforms in cascade was proposed: SIC, Power BI and Knime Analytics Platform. This combination responds to a series of needs of modern medicine and today's healthcare context due to the virus, allowing healthcare users to have BI tools available, such as dashboards with interactive graphics, data georeferencing, reports with the possibility of predefined templates, and the possibility of using artificial intelligence on the enormous amount of data that is produced by the territory due to the emergency.

3.6.2. SIC platform

The taking charge model is managed through the SIC platform which, in relation to the management of COVID-19 positive patients, aims to achieve the following objectives:

- Allow to assess the health condition in which a person who has had contact (tried or presumed) with the virus may be;
- Make the appropriate decisions regarding the need to perform a verification diagnosis swab;
- Managing the supply chain and the tracking both in the development phase and in the analysis phase of the swabs;

- Managing the follow-up of quarantined patients and eventually his treatment.

Patient management is divided within the system into different paths that originate from an assessment of the clinical conditions of the patient who actively participates from home, quickly providing information to the specialists involved who will thus be able to make the most appropriate decisions by limiting the movements of the subject.

Figure 12 shows the authentication screen for accessing the platform.

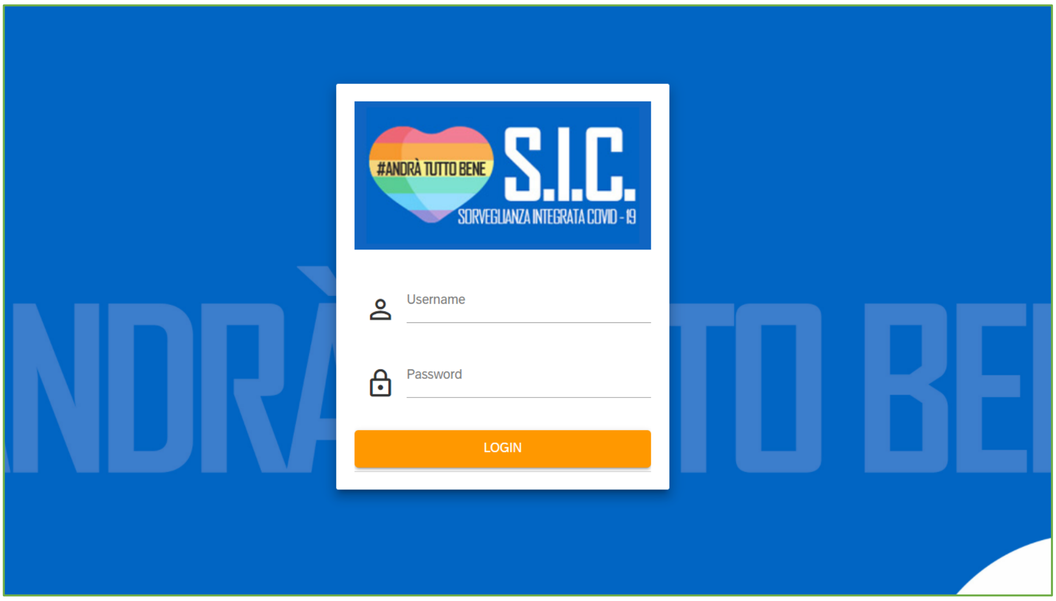


Figure 12. Authentication screen of SIC platform.

3.6.3. Main actors involved in the process

SIC platform is accessible from different profiles since it involves several actors. The profiles that can be identified and participating to the process are the following:

- MMG and PLS: belonging to the territorial network of the Local Districts of the ASL, they have the possibility to report new suspected cases and to manage their follow-up by collaborating with the U.S.C.A.

- U.S.C.A.: dedicated to the territorial surveillance of COVID-19 cases, they can:
    - Report a suspected case;
    - Manage the follow-up;
    - Take charge and handle the most fragile patients;
    - Evaluate the possibility of swabbing for both verification and control.
- Mobile Operating Units "Unità Mobili Operative": they are means equipped for the management of tampons at home. A special App dedicated to them has been created to allow the operators of the Mobile Operating Units to interact with the platform,
- Patients: they use the services of the platform through a dedicated App that allows them to answer NEWS 2 self-assessment questionnaires.

3.6.4. Technologic elements of the system

The information system consists of the following elements:

- Centre of Telematic Service "Centro Servizi Telematico" (CST): characterized by the presence of personnel dedicated to health surveillance (U.S.C.A.). The system summarizes the status of all managed patients, giving the possibility to filter and view the status of the single patient registered in the system;
- Web App for patients: it is used to allow interaction between the patient and the CST. The App allows users to deliver NEWS 2 evaluation questionnaires to patients who can respond directly from their smartphone;
- Web App for Mobile Operating Units: it is used to allow the interaction of Mobile Unit operators and the CST. The App allows users to confirm that the sample has been taken using a special swab

at a patient's home or at a place dedicated to screening. Patient-swab matching takes place by connecting the tax code with the unique identifier of the swab printed on the label with EAN code (Bar Code);

- DSS: it is used to perform the automatic evaluation of scores based on patient responses. Each type of questionnaire correctly completed by the patient is associated with colour values to establish the level of risk.

Figure 13 shows the main elements of the system.

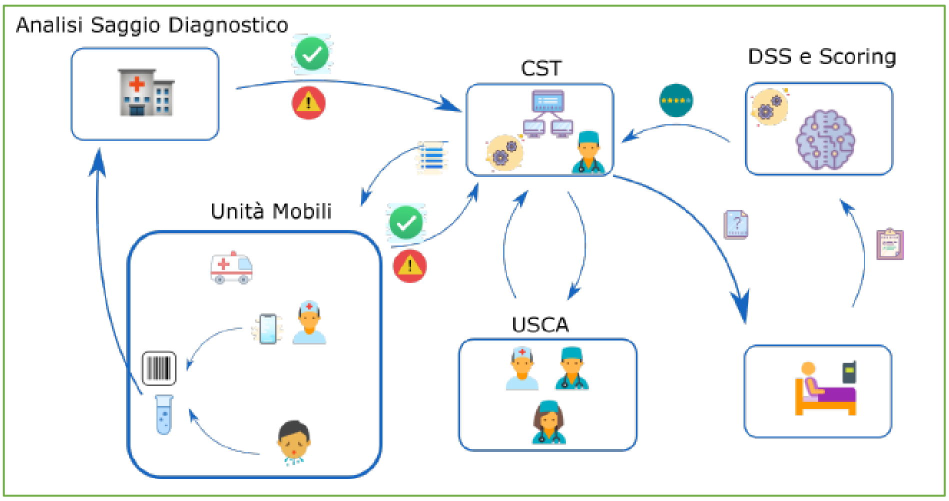


Figure 13. Technologic elements of the system SIC.

3.6.5. CST, MMG e PLS

The CST allows to have the full control and monitoring of the situation both from the point of view of patients at risk and from the administrative point of view. The CST is a system that allows multi-tenant access with different profiles and roles that can perform actions in relation to the permissions assigned to them. CST administrators and operators have full access rights and can:

- Register the Health Districts adhering to the territorial management protocol for Covid-19 patients;
- Visualize all the MMG related to a Health District;

- Manage the list of patients enrolled in the CST and allow for each of them to:
    - Read the personal information;
    - Insert and register the values of temperature, blood pressure and oxygen in the blood;
    - Assign a daily auto evaluation questionnaire;
    - Define the clinical picture (Pathologies, allergies, etc.).
- Manage the eventual alarms for suspected cases allowing also the enrolment;
- Define the requests of swabs for patients;
- Manage the Mobile Operational Units and assign to them patients who should receive a swab;
- Visualize the output of swabs for patients stored in the system;
- Visualize the reports of the swabs coming from the laboratories;

The CST platform makes the informative tools available to specialists, MMGs and possibly local authorities to report potential suspected positive patients. Downstream of the reports, it then allows the various preventive and assistance protocols to be initiated for patients reported as suspected positive. The CST also has supervision operators who, in addition to administering the platform, manage incoming calls from patients who, having not received feedback from other channels, spontaneously decide to report their suspected case. In this case, the patient is entered into the system and will be managed as we will see in the section dedicated to the patient.

The MMG belongs to a specific District and is logically associated with it within the CST information system. The MMG is responsible for active health surveillance carried out with telephone triage on his own initiative or by the patient who gets in touch with him. Based on the first telephone triage, the MMG may decide, if necessary, to communicate the name and address of the patients to the U.S.C.A. which assess the subjects with symptoms who must be considered as suspected COVID-19 cases or are COVID-19. The case of reporting by the MMG could originate from the following cases:

- A patient called the MMG due to a suspect of being positive;
- MMG suspected that a patient is positive due to the symptomatology.

Within SIC platform, the MMG or the PLS can open a new reporting case by entering into the system all the necessary clinical and medical information of the patient that can be useful to the U.S.C.A. doctors. At the same time, if patient's telephone number is entered, he is sent an invitation to run a questionnaire that allows users to calculate the NEWS 2 Score. The questionnaire can be filled out directly from the mobile phone thanks to the presence of the Web App.

3.6.6. Processes

Figure 14 shows the patient recruitment process which, as anticipated, can start from a self-report and/or a direct report from the U.S.C.A. or the MMG. In the last case, a telephone triage is carried out in order to allow the doctor to carry out an initial evaluation.

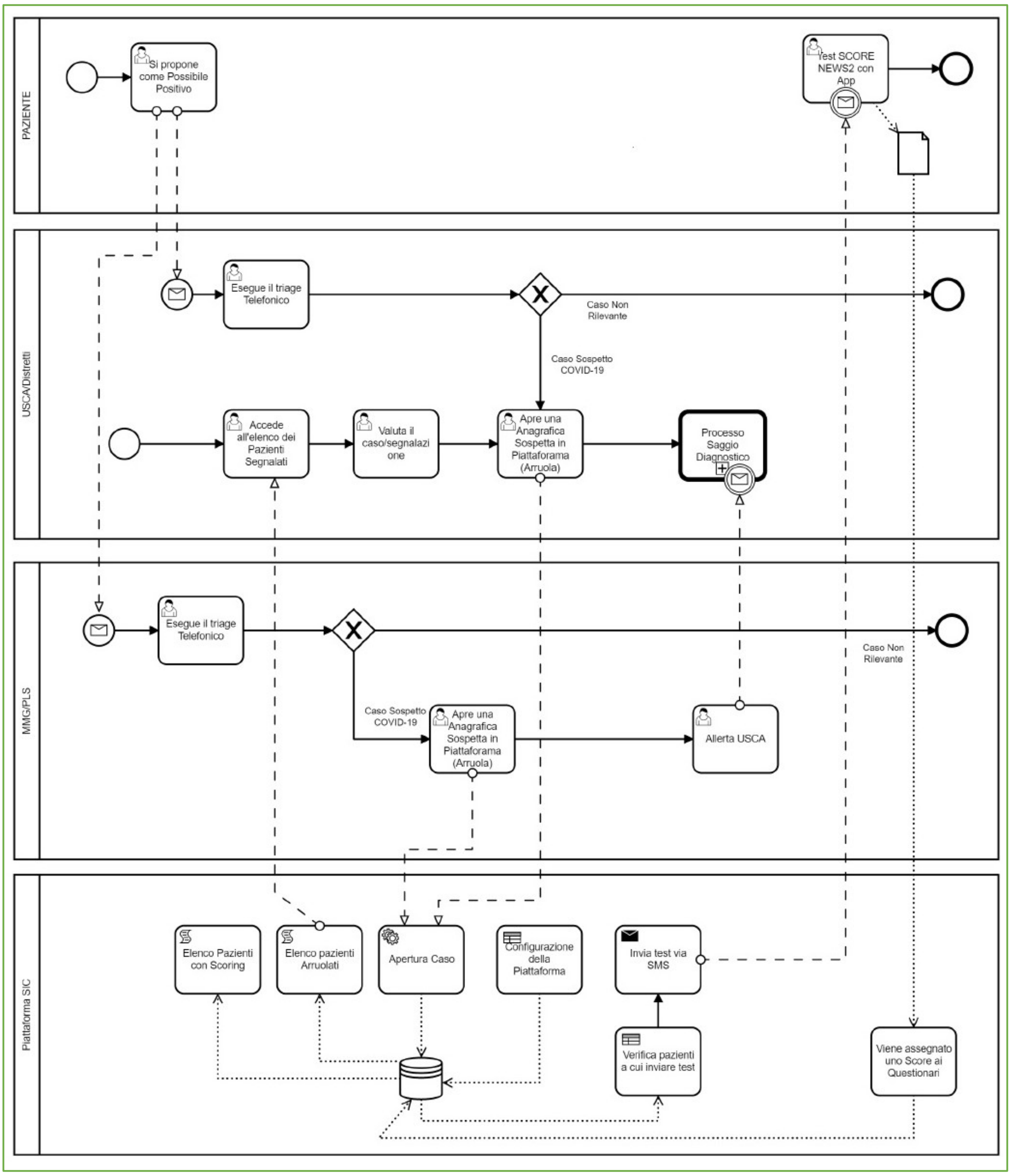

PAZIENTE
Si propone come Possibile Positivo
Test SCORE NEWS2 con App
USCA/Distretti
Esegue il triage Telefonico
Caso Non Rilevante
Caso Sospetto COVID-19
Accede all'elenco dei Pazienti Segnalati
Valuta il caso/segnalazione
Apre una Anagrafica Sospetta in Piattaforama (Arruola)
Processo Saggio Diagnostico
MMG/PLS
Esegue il triage Telefonico
Caso Non Rilevante
Caso Sospetto COVID-19
Apre una Anagrafica Sospetta in Piattaforama (Arruola)
Allerta USCA
Piattaforma SIC
Elenco Pazienti con Scoring
Elenco pazienti Arruolati
Apertura Caso
Configurazione della Piattaforma
Invia test via SMS
Verifica pazienti a cui inviare test
Viene assegnato uno Score ai Questionari

Figure 14. Recruitment process for COVID-19 patients.

When contacted via SMS, the patient receives a code to log into the platform and fill out a 10-point questionnaire:

1) Attendance of people;
2) Alerting attended places;

3) Verifying the presence of serious symptoms (breathing difficulties, altered consciousness, systolic and diastolic pressure, heart rate);
4) Presence of symptoms of COVID-19 (date);
5) Major symptoms (fever and cough);
6) Fever;
7) Quantifying fever (from 0 to 9);
8) Gastrointestinal symptoms;
9) Minor symptoms (tiredness, sore throat, g-lump, headache, muscle aches, nasal congestion, etc);
10) Epidemiologic link (exposure to ascertained or suspected cases, contacts with family members of suspected cases, attendance of healthcare environments with ascertained or suspected cases).

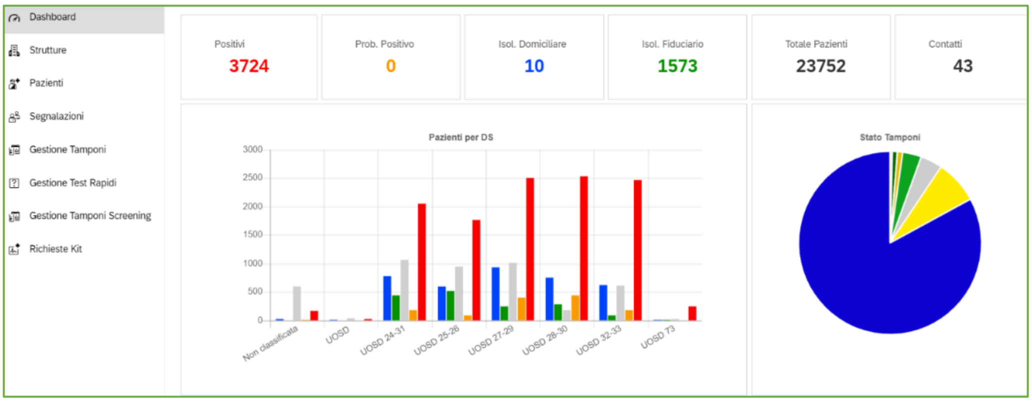
Alternatively, the patient can be engaged via telephone and is subjected to a questionnaire similar to the one found on the platform.

3.6.7. Web interfaces in SIC platform

The first interface encountered on the platform is the one in figure 15, it is an interactive BI dashboard that allows users to view:

- The number of positive patients,
- The overall number of managed patients,
- The number of quarantined people,
- The number of suspected cases.

In addition, there is a simple bar chart showing the breakdown by Simple Departmental Operating Unit (U.O.S.D.).



Positivi	Prob. Positivo	Isol. Domiciliare	Isol. Fiduciario	Totale Pazienti	Contatti
3724	0	10	1573	23752	43

Figure 15. Basic interactive dashboard integrated in SIC platform.

The patient management interface is the one in figure 16. The following buttons are present:

- Reservation: it allows user to view requests already sent, it is possible to search for patients using the filters provided in the appropriate section.
- Requests: it allows user to insert within the platform the candidates receiving a swab.
- Assigned: it shows the summary of patients who received the unit where they will physically be subjected to the swab.
- Executed: It shows an overview of patients already subjected to the swab.
- Not executed: in addition to the “executed” function, it is possible to check through this function any not subjected to swab.
- Archive: it shows all patients in the database.

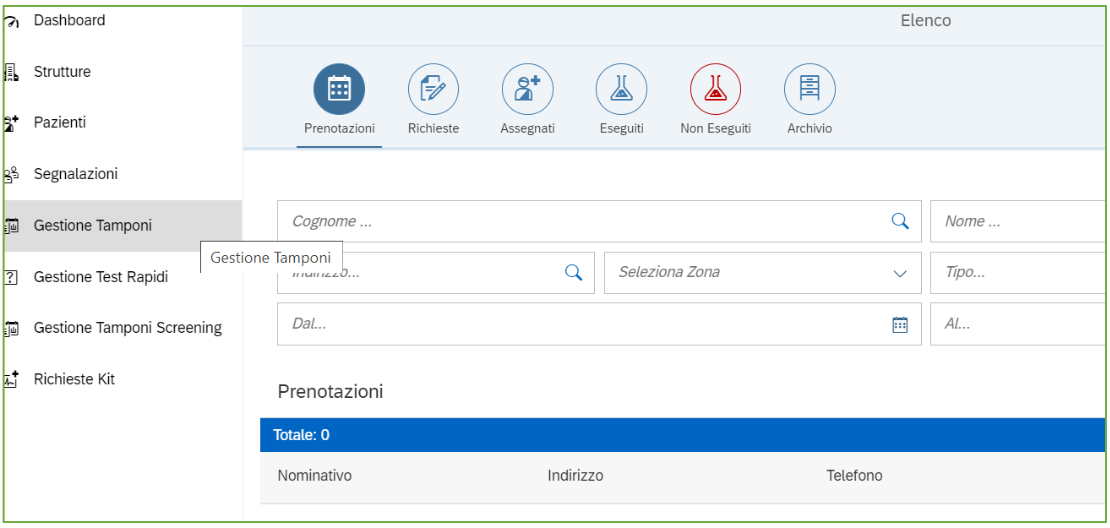


Figure 16. Swab management interface in SIC platform.

3.6.8. The integration of SIC with Power BI

The workflow to integrate SIC with Power BI and, possibly, with Knime Analytics Platform, is represented in figure 17

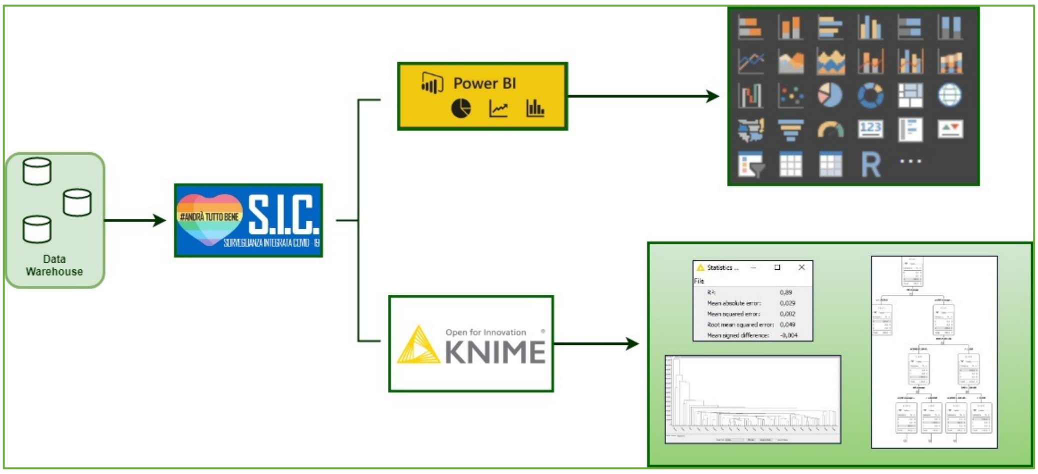


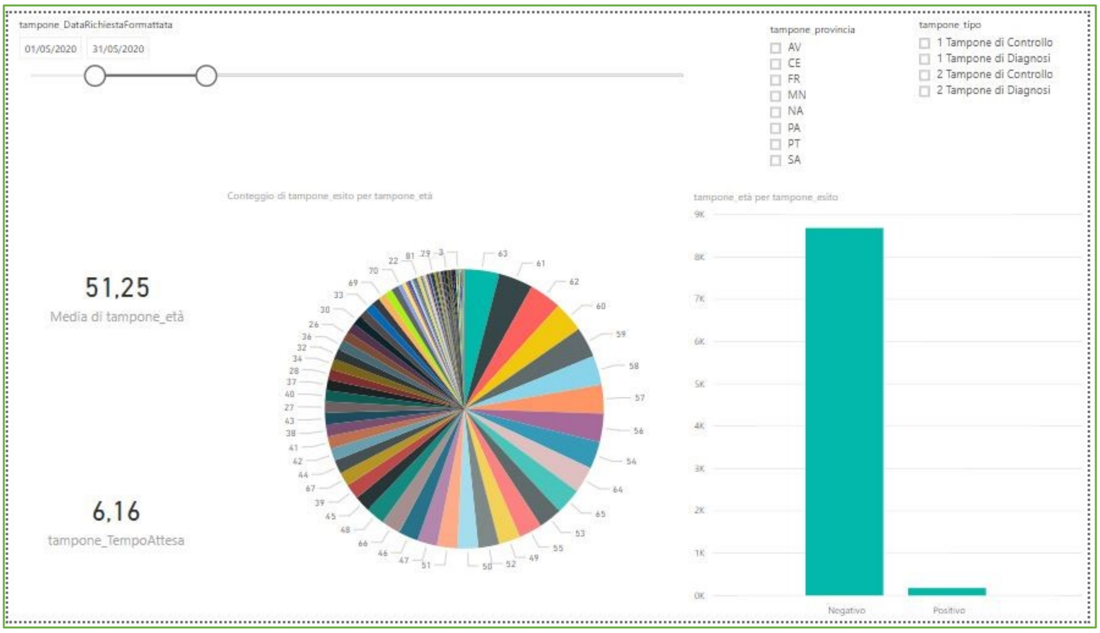
Figure 17. Workflow of the system made up of a DW, SIC platform and analytical tools of Power BI and Knime analytics platform.

The data of 50271 patients related to the management of swabs were collected in a completely anonymized way in order to link the file containing the data to Power BI. For privacy reasons, it was preferred to avoid connecting the tool directly to the DW built and managed via DBMS. Before going on, I need to mention that there is no intention in this thesis to perform an epidemiological analysis on the data. The data will be used only as a test for the platform.

The following variables were collected:

- Date and time of the request for swab;
- Date and time of the receipt of the swab results;
- Type of swab (check or diagnostic);
- Output of the swab;
- City where patient lives;
- Age of the patient.

The report that could help the management control of the ASL is represented in figure 18.



tampone_DataRichiestaFormattata
01/05/2020 31/05/2020

tampone_provincia
□ AV
□ CE
□ FR
□ MN
□ NA
□ PA
□ PT
□ SA

tampone_tipo
□ 1 Tampone di Controllo
□ 1 Tampone di Diagnosi
□ 2 Tampone di Controllo
□ 2 Tampone di Diagnosi

51,25
Media di tampone_età

6,16
tampone_TempoAttesa

Figure 18. Report of Power BI including filters, labels, pie and bar charts.

This interactive dashboard contains a series of filters at the top:

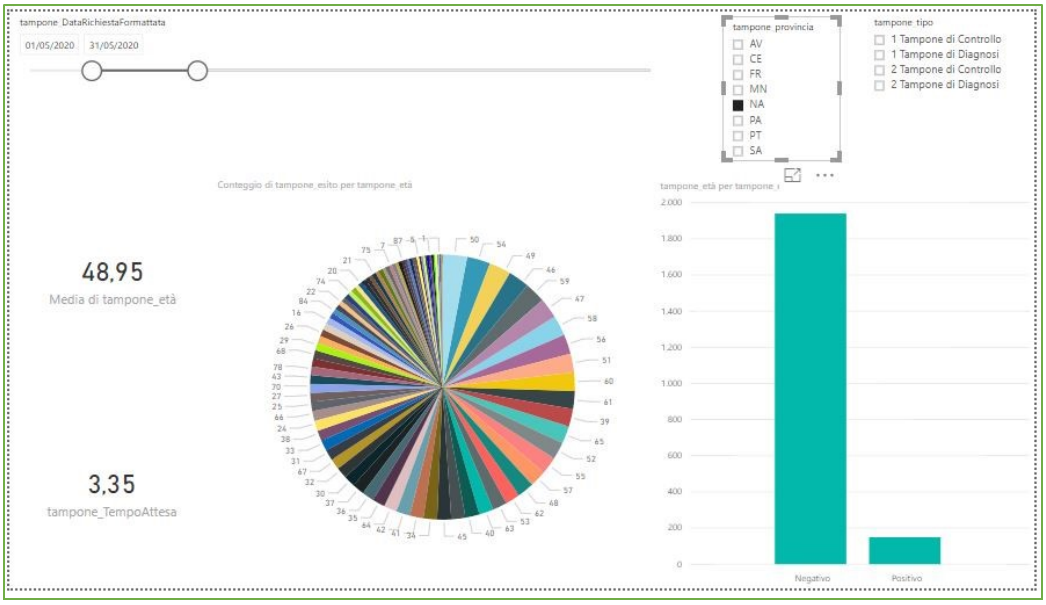
- At the top left, there is a time filter that, by dragging the cue balls, allows the users to choose the time period for the representation of the data;
- At the top centre, there is the possibility to filter by province;

- At the top right, the type of swab to represent can be chosen.

Furthermore, there are some graphs and numerical information:

- At the bottom left, there are the average waiting time for the swab result, measured in days, and the average age of the patients;
- In the lower centre, there is a pie chart representing the number of patients who received a swab in relation to age;
- In the lower left centre, there is a bar plot representing the quantity of negative and positive swabs.

The interactivity of the dashboard is particularly useful as shown in figures 19, 20, 21.



tampone DataRichiestaFormattata

01/05/2020 31/05/2020

tampone provincia

□ AV
□ CE
□ FR
□ MN
■ NA
□ PA
□ PT
□ SA

tampone tipo

□ 1 Tampone di Controllo
□ 1 Tampone di Diagnosi
□ 2 Tampone di Controllo
□ 2 Tampone di Diagnosi

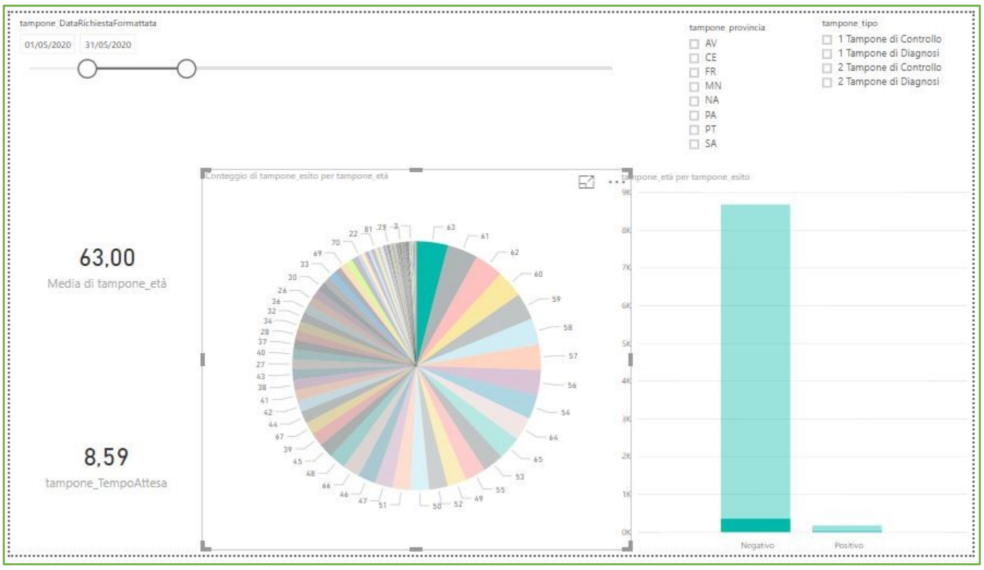
48,95

Media di tampone_età

3,35

tampone_TempoAttesa

Figure 19. Report of Power BI after updating the province filter by choosing Naples "NA".



tampone_DataRichiestaFormattata

01/05/2020 31/05/2020

tampone_provincia

□ AV
□ CE
□ FR
□ MN
□ NA
□ PA
□ PT
□ SA

tampone_tipo

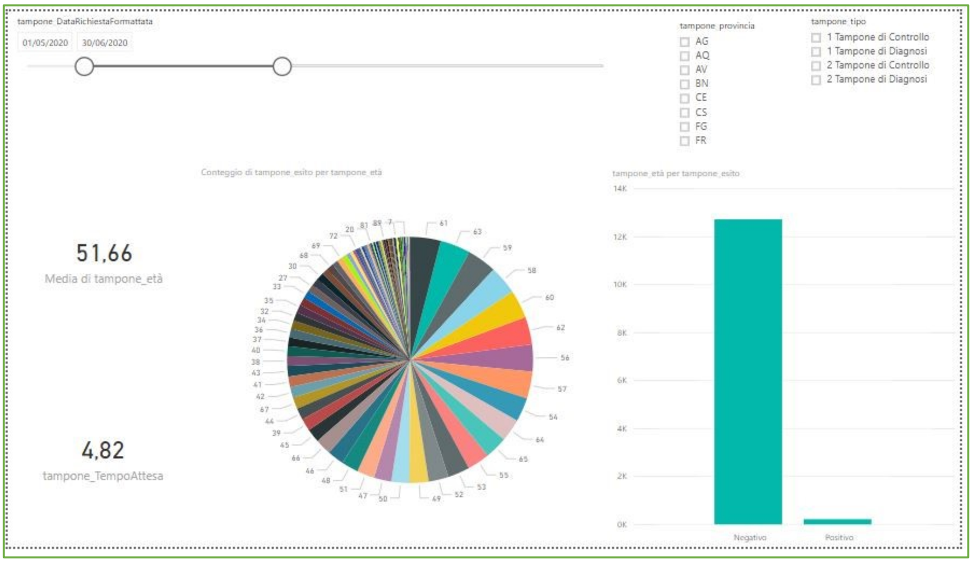
□ 1 Tampone di Controllo
□ 1 Tampone di Diagnosi
□ 2 Tampone di Controllo
□ 2 Tampone di Diagnosi

63,00
Media di tampone_età

8,59
tampone_TempoAttesa

Conteggio di tampone_esito per tampone_età
tampone_età per tampone_esito
9K
8K
7K
6K
5K
4K
3K
2K
1K
0K
Negativo
Positivo

Figure 20. Report of Power BI updated with age filter on 63 years.



tampone_DataRichiestaFormattata

01/05/2020 30/06/2020

tampone_provincia

□ AG
□ AQ
□ AV
□ BN
□ CE
□ CS
□ FG
□ FR

tampone_tipo

□ 1 Tampone di Controllo
□ 1 Tampone di Diagnosi
□ 2 Tampone di Controllo
□ 2 Tampone di Diagnosi

51,66
Media di tampone_età

4,82
tampone_TempoAttesa

Conteggio di tampone_esito per tampone_età
tampone_età per tampone_esito
14K
12K
10K
8K
6K
4K
2K
0K
Negativo
Positivo

Figure 21. Report of Power BI after updating the time filter.

In figure 19, the report has been updated, compared to figure 18, by clicking on "NA" in the province filter; therefore, the dashboard was automatically updated, showing, in the period defined by the time filter, the average values of age and waiting time for the outcome, the pie chart for patients in the province of Naples and the positives / negative in the province of Naples

In figure 20, the report has been updated, compared to figure 18, by clicking on the slice of the pie chart with patients aged 63 years; therefore, the dashboard is automatically updated, showing, in the period defined by the time filter, the average values of age (clearly 63 years) and waiting time for the outcome, the pie chart that highlights the selected slice and in the bar plot the positives / negatives of 63 years.

In figure 21, the report has been updated, compared to figure 18, extending the time period from 1 to 2 months. All the other graphs have automatically updated their values, in the bar plot the reference is no longer “9k” but “14k” to testify the increase in the number of patients considered.

4. Discussion

The use of ICT tools has been growing intensely in recent years and it is possible to see their practical application in everyday life. This application can be particularly useful if implemented not only in research, where it is already used (in a different way) on a large scale as seen in the international literature, but also at an industrial level.

Some tangible examples are present. Deepmind is an English company, owned by Alphabet (a US company which also owns Google), and has been involved for years in the field of artificial intelligence. Among its divisions there was the "Health section", whose goal was to insert Artificial Intelligence in healthcare: applications, clouds and dedicated solutions. Since November 2019 Deepmind Health has merged with Google Health. The Moorfields Eye Hospital NHS Foundation Trust, a famous London eye hospital, collaborated with Deepmind in 2016 to explore and implement solutions that can make researchers explore causes that lead to vision loss: diabetic retinopathy and age-related macular degeneration. The goal was to analyse medical images using machine learning algorithms in order to try to early diagnose the disease. Similarly, Deepmind collaborated with University College London Hospital to use Artificial Intelligence technologies to help planning radiotherapy treatment for patients with head and neck cancer by analysing CT images.

The benefits that the use of these technologies can bring to health are countless [45-46]:

- Healthcare risk assessment;
- Real-time estimation of health risks for the community, predicting problems and suggesting solutions;
- Carrying out repetitive work, such as the analysis of tests, X-rays, computed tomography or the entry of data into the computer system;

- Reduction of diagnostic and therapeutic errors, inevitable in clinical practice carried out by healthcare personnel;
- Management of medical records and performance analysis of a single health facility;
- Development of "precision" medicine with the implementation of new drugs based on faster processing of the mutations of the analysed disease;
- Digital consultancy and health monitoring services, operating as a "digital nurse".

Similarly, there are also many advantages that can be brought by a BI solution at a healthcare level as hospitals, in a certain sense, are no different from all the other companies because they have to pay attention to income, costs, use of resources and quality of services [47]; many of the advantages provided by analytics systems have already been extensively discussed in the literature [48-50].

The solution proposed in this thesis is open source and absolutely user-friendly; to start the architectural layout is shown in figure 22.

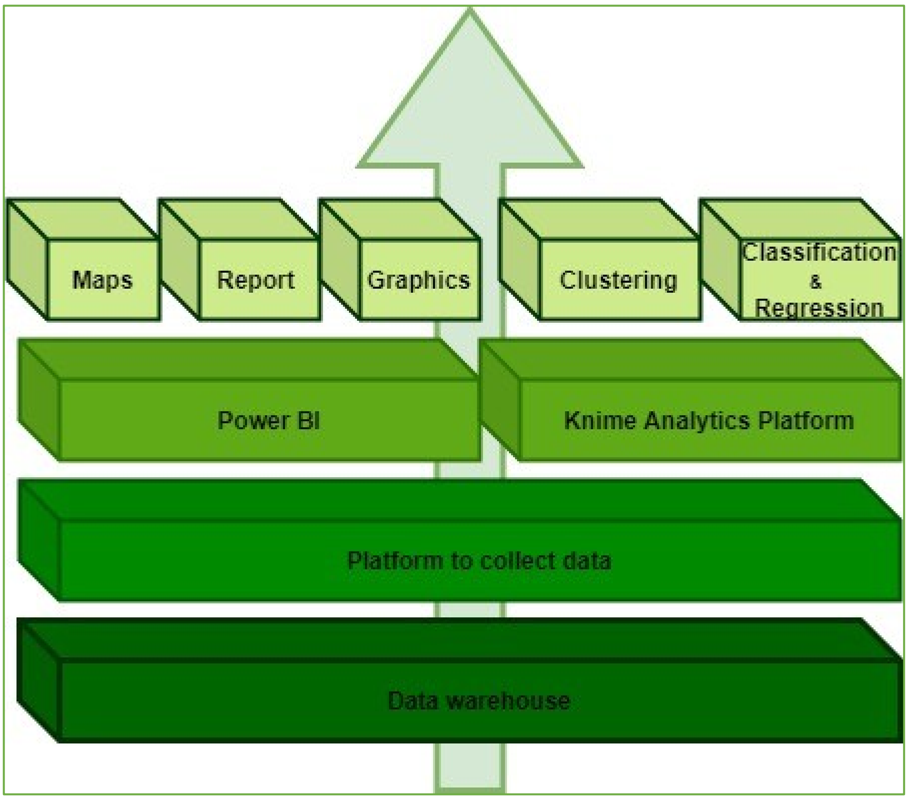


Figure 22. Architectural system of the proposed ICT solution.

As shown in figure 22, the proposed system involves the presence of a DW at the base therefore, upstream of the solution, there is always a source of structured data that can be managed through DBMS. This has to be considered the starting point and also the foundation of any technological solution aimed at analysing data; indeed, it is impossible to think of making decisions based on data if these have not been arranged in an organized and usable way for the available technologies.

The next layer of the architecture is also fundamental since an adequate data collection system must simplify the life of both the users involved in the data collection and the data scientists who are downstream of the collection process. This must be done in a systematic and accurate way, trying both to insert the appropriate controls to avoid human errors that, clearly, can happen, and entering data in a syntactically or semantically incorrect way (for example, inserting in the field "city" the street in which the patients live makes it difficult and onerous to georeference the data)

The two platforms are positioned at even higher levels, Power BI to make maps, graphs and reports and Knime Analytics platform to carry out classification and regression studies (but also unsupervised learning if desired). As shown in the practical application of both tools, they have allowed the creation of powerful and scalable data models, the analysis and prediction of the incidence of diseases and potentially the identification of the causes related to suboptimal economic management. They make it possible to acquire and process data from heterogeneous sources, as it is possible to connect platforms to even different data sources, interactively through dashboards and machine learning workflows. Among the various capabilities of the combination of tools there are the possibilities of:

- Analysing and aggregating data using a DW, designed and implemented appropriately;
- Extracting the hidden knowledge in the data, identifying patterns of regularity, with the application of:
  - Statistical techniques;
  - Predictive models,
  - Advanced data mining algorithms,
- Enable the route from data to information and, thus, to knowledge;
- Classifying, identifying and interpreting data related to specific patients,
- Overall improvement of the care and management process;
- Allowing operators in the health sector to move from a vertical to a horizontal approach, favouring the creation of a reality in which all the elements involved in the process are

connected and there is a strong collaboration between the various actors operating in the operative sector.

The main repercussions deriving from the use of the proposed solution in the health sector are:

- Building a shared knowledge of the data;
- Transforming historical data into a valid decision-support tool;
- Reducing space/time obstacles in the storage, management and processing of information;
- Optimizing the management and analysis of the data produced by the various units of the hospital;
- Determining the quantification of the volumes of activity by service, divided into sub-periods with respect to a reference one;
- Pursuing a flexible balance between people's health needs and the economic and financial sustainability of the entire system.

In addition, there are various needs in the health sector that find an answer in this technological solution; it starts from the management of large amounts of heterogeneous data that are produced at hospital level up to the design of reports and interactive dashboards that help strategic business management in decision-making processes, passing through the help to clinicians in the prognosis and diagnosis phase through forecasting, pattern recognition and statistical analysis.

4.1. Conclusion, limitations and future development

The conducted research provided users with user-friendly and open source tools for the implementation of machine learning and the use of BI within clinical practice. However, except for the SIC platform, which is currently in use at ASL of Naples, the other technologies have not been introduced into clinical routine. An economic study for cost-benefit, cost-utility and cost-effectiveness evaluation could be the last step to be carried out before then testing these ICT technologies within the hospital and outpatient clinical routine.

In order to take this step forward, it is necessary to involve all the actors of the health process together with the data scientists who can implement these technologies. In fact, there are researchers and clinical figures who have not yet become familiar with these practices also due to results that are sometimes not reproducible or not clearly presented [51].

There are already guidelines for drafting reports of predictive models, some examples are the "Strengthening the Reporting of Observational Studies in Epidemiology" (STROBE) and the "Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis" (TRIPOD) which are, however, focused on observational studies or on the development of multivariate predictive models without mentioning machine learning analyses. It would probably be helpful to define a standard model to draw up also machine learning analysis reports where there are clear requirements to deem a satisfactory and reliable analysis. Following, there are the general features that a ML analysis should have in order to be fully considered in clinical research: study population, repeatability, validation, results and comparison with the human.

On the study population there is not a real requirement but there are suggestions in literature which are important in order to avoid the “dimensionality curse”. The number of subjects should be 10 times the number of variables that will be included in the model.

Another requirement, related to study population, is the type of analysed pathology: on the one hand, if the analysis has as a target a common pathology, it is easier to obtain several subjects and also to design a balanced dataset (comparable number of healthy and affected patients). On the other hand, there is a different situation when analysing a rare pathology; indeed, acquiring a satisfactory dataset is more difficult and probably a smaller dataset should be accepted by the research community in order to acquire new knowledge (thus, being flexible regarding the study population requirement).

As regards the repeatability, other researchers should be able to repeat the analyses shown in a research paper; therefore, the authors should explain precisely how the analysis was performed: tools, algorithms, pre-processing of the data, evaluation metrics.

The validation is a strict requirement: ML analysis needs a training and a test set but there are several methods to validate the results. CV gives the chance to obtain more reliable results also when an external dataset to validate the models is not available [18]:

- Hold-out should be used only with large datasets (N>1000).
- Leave one out CV should be used with small datasets (N<100).
- K-fold CV with k=10 (as suggested by Kohavi et al. [18]) could be used in the other cases (100<N<1000).

In the Results section, there should be the use of the appropriate metrics that allow readers to understand the real value of the implemented models. Accuracy, sensitivity, specificity and Area Under the Curve Receiver Operating Characteristics (known as AUCROC) should always be used. Moreover, in a ML analysis the features importance should always be considered since it allows together analysts and clinicians to acquire knowledge on a feature/variable which had not been considered before.

Finally, the comparison between the human and the ML performance, excluding in the feasibility studies, should be comparable to consider successful the presented analysis. Indeed, the aim should always be adding value to the clinical practice when introducing a ML-based tool.

There are still many questions that have still not received an answer regarding the actual introduction of machine learning techniques within the healthcare sector during clinical practice [52] and they will become more and more in the following years. In the meantime, such techniques become more robust and the international literature continues to show their usefulness in a myriad of application fields.

Acronyms

ADA-B: Ada-boosting

ASL: Aziende Sanitarie Locali

AUCROC: Area Under the Curve Receiver Operating Characteristics

BI: Business Intelligence

CAD: Coronary Artery Disease

CHD: Coronary heart disease

CHF: Chronic heart failure

Co.S.A.: “Correlazione Salute Ambiente”

COG: Cognitive dual task of gait analysis

CST: Centro Servizi Telematici

CVD: Cardiovascular disease

CZT: Cadmium–zinc–telluride

DBMS: Database Management System

DW: Data Warehouse

DSS: Decision Support Systems

DRG: Diagnosis Related Group

DT: Decision Tree

FHR: Fetal Heart Rate

GAIT: Single task of gait analysis

GB: Gradient boosting tree

ICT: Information and communication technology

KD: Knowledge Discovery

KNN: K Nearest Neighbour

LDA: Linear Discriminant Analysis

MCI: Mild Cognitive Impairment

MLP: Multilayer Perceptron

MMG: Medico di Medicina Generale

MOT: Motor Dual task of gait analysis

MPI: Myocardial Perfusion Imaging

NB: Naïve Bayes

NTRA: Nonlinear Trimodal Regression Analysis

PCA: Principal Component Analysis

PD: Parkinson’s Disease

PLS: Pediatra di Libera Scelta

PSP: Progressive Supranuclear Paralysis

RF: Random Forests

SDp: Variability of eye-positioning

SIC: “Sorveglianza Integrata Covid-19”

STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

SVM: Support Vector Machine

TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis

U.O.S.D.: Unità Operativa Semplice Dipartimentale

U.S.C.A.: Unità Speciali di Continuità Assistenziale

VA: Visual acuity

Index of figures

Index of tables

References

1. Isik, O., Jones, M. C., & Sidorova, A. (2011). Business intelligence (BI) success and the role of BI capabilities. Intelligent systems in accounting, finance and management, 18(4), 161-176.
2. Bose, R. (2003). Knowledge management-enabled health care management systems: capabilities, infrastructure, and decision-support. Expert systems with Applications, 24(1), 59-71.
3. Olszak, C. M., & Batko, K. (2012, September). The use of business intelligence systems in healthcare organizations in Poland. In 2012 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 969-976). IEEE.
4. Foshay, N., & Kuziemsky, C. (2014). Towards an implementation framework for business intelligence in healthcare. International Journal of Information Management, 34(1), 20-27.
5. Mettler, T., & Vimarlund, V. (2009). Understanding business intelligence in the context of healthcare. Health informatics journal, 15(3), 254-264.
6. Hunt, D. L., Haynes, R. B., Hanna, S. E., & Smith, K. (1998). Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. Jama, 280(15), 1339-1346.
7. Efron, B. (1990). More efficient bootstrap computations. Journal of the American statistical association, 85(409), 79-89
8. Binder, K., Heermann, D., Roelofs, L., Mallinckrodt, A. J., & McKay, S. (1993). Monte Carlo simulation in statistical physics. Computers in Physics, 7(2), 156-157
9. Geyer, C. J. (1992). Practical markov chain monte carlo. Statistical science, 473-483.

10. Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications–A decade review from 2000 to 2011. Expert systems with applications, 39(12), 11303-11311

11. Wiens, J., & Shenoy, E. S. (2018). Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clinical Infectious Diseases, 66(1), 149-153.

12. Luhn, H. P. (1958). A business intelligence system. IBM Journal of research and development, 2(4), 314-319.

13. Cleland, D. I., & King, W. R. (1975). Competitive business intelligence systems. Business Horizons, 18(6), 19-28.

14. Ranjan, J. (2009). Business intelligence: Concepts, components, techniques and benefits. Journal of Theoretical and Applied Information Technology, 9(1), 60-70.

15. Gangadharan, G. R., & Swami, S. N. (2004, June). Business intelligence systems: design and implementation strategies. In 26th International Conference on Information Technology Interfaces, 2004. (pp. 139-144). IEEE.

16. Knight, D., Knight, B., Pearson, M., Quintana, M., & Powell, B. (2018). Microsoft Power BI complete reference: bring your data to life with the powerful features of Microsoft Power BI. Packt Publishing Ltd.

17. Love, B. C. (2002). Comparing supervised and unsupervised category learning. Psychonomic bulletin & review, 9(4), 829-835.

18. Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).

19. Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1-2), 273-324.

20. Lewis D.D. (1998) Naive (Bayes) at forty: The independence assumption in information retrieval. In: Nédellec C., Rouveirol C. (eds) Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science

(Lecture Notes in Artificial Intelligence), vol 1398. Springer, Berlin, Heidelberg. https://doi.org/10.1007/BFb0026666

21. Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics, 21(3), 660-674.
22. Kataria, A., & Singh, M. D. (2013). A review of data classification using k-nearest neighbour algorithm. International Journal of Emerging Technology and Advanced Engineering, 3(6), 354-360
23. Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. Neural processing letters, 9(3), 293-300
24. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
25. Rätsch, G., Onoda, T., & Müller, K. R. (2001). Soft margins for AdaBoost. Machine learning, 42(3), 287-320.
26. Friedman, J. H. (2002). Stochastic gradient boosting. Computational statistics & data analysis, 38(4), 367-378.
27. Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., ... & Wiswedel, B. (2009). KNIME-the Konstanz information miner: version 2.0 and beyond. AcM SIGKDD explorations Newsletter, 11(1), 26-31.
28. Fillbrunn, A., Dietz, C., Pfeuffer, J., Rahn, R., Landrum, G. A., & Berthold, M. R. (2017). KNIME for reproducible cross-domain analysis of life science data. Journal of biotechnology, 261, 149-156.
29. Warr A. W. "Scientific workflow systems: Pipeline Pilot and KNIME," Journal of Computer-Aided Molecular Design, vol. 26, pp. 801-804, 27 May 2012.
30. Ricciardi, C., Cantoni, V., Improta, G., Iuppariello, L., Latessa, I., Cesarelli, M., ... & Cuocolo, A. (2020). Application of data mining in a cohort of Italian subjects undergoing myocardial perfusion imaging at an academic medical center. Computer Methods and Programs in Biomedicine, 189, 105343.

31. Mannarino, T., Assante, R., Ricciardi, C., Zampella, E., Nappi, C., Gaudieri, V., ... & Cuocolo, A. (2019). Head-to-head comparison of diagnostic accuracy of stress-only myocardial perfusion imaging with conventional and cadmium-zinc telluride single-photon emission computed tomography in women with suspected coronary artery disease. Journal of Nuclear Cardiology, 1-10

32. Cantoni, V., Green, R., Ricciardi, C., Assante, R., Zampella, E., Nappi, C., ... & Giordano, A. (2020). A machine learning-based approach to directly compare the diagnostic accuracy of myocardial perfusion imaging by conventional and cadmium-zinc telluride SPECT. Journal of nuclear cardiology: official publication of the American Society of Nuclear Cardiology.

33. Ricciardi, C., Edmunds, K.J., Recenti, M. et al. Assessing cardiovascular risks from a mid-thigh CT image: a tree-based machine learning approach using radiodensitometric distributions. Sci Rep 10, 2863 (2020).

34. Ricciardi, C., Valente, A. S., Edmund, K., Cantoni, V., Green, R., Fiorillo, A., Picone, I., Santini, S., & Cesarelli, M. (2020). Linear discriminant analysis and principal component analysis to predict coronary artery disease. Health Informatics Journal, 26(3), 2181–2192. https://doi.org/10.1177/1460458219899210

35. Harris, T. B. et al. Age, gene/environment susceptibility–Reykjavik study: multidisciplinary applied phenomics. Am. J. Epidemiol. 165(9), 1076–87 (2007).

36. Davis III, R. B., Ounpuu, S., Tyburski, D., & Gage, J. R. (1991). A gait analysis data collection and reduction technique. Human movement science, 10(5), 575-587.

37. Ricciardi, C., Amboni, M., De Santis, C., Improta, G., Volpe, G., Iuppariello, L., ... & Cesarelli, M. (2019). Using gait analysis' parameters to classify Parkinsonism: A data mining approach. Computer methods and programs in biomedicine, 180, 105033.

38. Ricciardi, C., Amboni, M., De Santis, C., Ricciardelli, G., Improta, G., D'Addio, G., ... & Cesarelli, M. (2020, June). Machine learning can detect the presence of Mild cognitive impairment in patients affected by Parkinson's Disease. In 2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA) (pp. 1-6). IEEE.

39. Ricciardi, C., Amboni, M., De Santis, C., Ricciardelli, G., Improta, G., Cesarelli, G., ... & Barone, P. (2020, June). Classifying patients affected by Parkinson's disease into freezers or non-freezers through machine learning. In 2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA) (pp. 1-6). IEEE.

40. Romeo, V., Ricciardi, C., Cuocolo, R., Stanzione, A., Verde, F., Sarno, L., ... & Maurea, S. (2019). Machine learning analysis of MRI-derived texture features to predict placenta accreta spectrum in patients with placenta previa. Magnetic resonance imaging, 64, 71-76.

41. Romeo, V., Cuocolo, R., Ricciardi, C., Ugga, L., Cocozza, S., Verde, F., ... & Elefante, A. (2020). Prediction of tumor grade and nodal status in oropharyngeal and oral cavity squamous-cell carcinoma using a radiomic approach. Anticancer Research, 40(1), 271-280.

42. Stanzione, A., Ricciardi, C., Cuocolo, R., Romeo, V., Petrone, J., Sarnataro, M., ... & Brunetti, A. (2020). MRI Radiomics for the Prediction of Fuhrman Grade in Clear Cell Renal Cell Carcinoma: a Machine Learning Exploratory Study. Journal of Digital Imaging, 1-9.

43. Ricciardi, C., Improta, G., Amato, F., Cesarelli, G., & Romano, M. (2020). Classifying the type of delivery from cardiotocographic signals: A machine learning approach. Computer Methods and Programs in Biomedicine, 196, 105712.

44. Improta, G., Ricciardi, C., Cesarelli, G., D'Addio, G., Bifulco, P., & Cesarelli, M. (2020). Machine learning models for the prediction of acuity and variability of eye-positioning using features extracted from oculography. Health and Technology, 10(4), 961-968.

45. Wiens, J., & Shenoy, E. S. (2018). Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clinical Infectious Diseases, 66(1), 149-153.

46. Bogdanova, A., Attoh-Okine, N., & Sakurai, T. (2020). Risk and Advantages of Federated Learning for Health Care Data Collaboration. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, 6(3), 04020031.

47. Ashrafi, N., Kelleher, L., & Kuilboer, J. P. (2014). The impact of business intelligence on healthcare delivery in the USA. Interdisciplinary Journal of Information, Knowledge, and Management, 9, 117-130.

48. Ivan, M., & Velicanu, M. (2015). Healthcare industry improvement with business intelligence. Informatica Economica, 19(2), 81.

49. Isazad Mashinchi, M., Ojo, A., & Sullivan, F. J. (2019, January). Analysis of Business Intelligence Applications in Healthcare Organizations. In Proceedings of the 52nd Hawaii International Conference on System Sciences.

50. Lee, S. Y. (2018, March). Architecture for business intelligence in the healthcare sector. In IOP Conf Ser Mater Sci Eng (Vol. 317).

51. Singh, K., Beam, A. L., & Nallamothu, B. K. (2020). Machine Learning in Clinical Journals: Moving From Inscrutable to Informative. Circulation: Cardiovascular Quality and Outcomes 13 (10) https://doi.org/10.1161/CIRCOUTCOMES.120.007491

52. Vollmer, S., Mateen, B. A., Bohner, G., Király, F. J., Ghani, R., Jonsson, P., ... & Granger, D. (2018). Machine learning and AI research for patient benefit: 20 critical questions on transparency, replicability, ethics and effectiveness. arXiv preprint arXiv:1812.10404.

Acknowledgement

Il percorso di dottorato termina con questa tesi, è stato un percorso di tre anni molto intenso ed allo stesso tempo molto stimolante ed interessante. Ci sono diverse persone che sento di dover ringraziare (in ordine alfabetico).

Prof. Mario Cesarelli, mio supervisore accademico, punto di riferimento e guida durante il percorso di dottorato a cui devo un grande grazie per le opportunità di crescita e gli innumerevoli spunti di saggezza che mi ha sempre fornito.

Prof. Paolo Gargiulo, mio supervisore nell'esperienza estera, uomo da enormi doti umane che mi ha stimolato accademicamente, supportato ed aiutato a vivere un'esperienza professionale e di vita di altissimo profilo.

Prof. Giovanni Improta, con cui ho avuto il piacere di collaborare in maniera proficua e continua per tutto il dottorato, devo ringraziarlo per le tante opportunità di ricerca e per gli spunti di crescita che mi ha fornito.

Dott. Pasquale Russo e dott. Mario Alessandro Russo, mi hanno dato l'opportunità di vivere un'esperienza aziendale con la Gesan s.r.l. (San Nicola la Strada, (CE)), so che ci legherà reciproca stima.

Un ringraziamento va anche al dipartimento di Scienze Biomediche Avanzate ed a tutti i suoi componenti, partendo dal Coordinatore, prof. Alberto Cuocolo, ed includendo tutti i professori, ricercatori, assegnisti e dottorandi per avermi permesso di integrarmi e collaborare in maniera sempre fruttifera.

Ringrazio poi la mia famiglia, da cui ho avuto da sempre supporto ed affetto incondizionato nel mio percorso di crescita, Riccardo, Susy e Claudia che mi hanno dato una seconda famiglia in questi ultimi tre anni.

Ringrazio i miei amici storici di essere ancora presenti dopo tanti anni, non li nomino ma ognuno di loro è consapevole dell'affetto che ci lega.

Infine, ringrazio, in poche parole, per avermi sempre sostenuto con fervore e con amore, Debora, la mia migliore amica, la mia partner di vita e la madre di mia figlia Aurora, il mio nuovo piccolo grande tesoro.

UNIONE EUROPEA
Fondo Sociale Europeo

Ministero dell'Università
e della Ricerca

PON
RICERCA
E INNOVAZIONE
2014 - 2020

Dottorando	Carlo Ricciardi
Tutor	Mario Cesarelli
Coordinatore	Alberto Cuocolo
Corso di Dottorato	Scienze Biomorfologiche e Chirurgiche
Ciclo	XXXIII
Codice borsa e n.	DOT1318209 – Borsa 4
CUP	E62G17000000006
Titolo Progetto	Progettazione, implementazione e realizzazione di una piattaforma integrata destinata all'e-Public Health, per l'analisi di dati sanitari ed il supporto al controllo di gestione nelle aziende sanitarie.

La borsa di dottorato è stata cofinanziata con risorse del

Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 (CCI 2014IT16M2OP005),

Fondo Sociale Europeo, Azione I.1 "Dottorati Innovativi con caratterizzazione Industriale"