POLITECNICO DI TORINO

Faculty of Engineering
Bachelor of Science in Mechatronics Engineering

Master's Degree Final Paper

Design of an IoT platform for predictive maintenance in spot welding

Candidato: Camperi Daniele

Relatori: Prof.ssa Bruno Giulia; Prof. De Maddis Manuela; Prof. Russo Spena Pasquale

A.A. 2020/2021

Index

Abstract

With the advent of Industry 4.0, Smart Manufacturing technology is implemented in industry sector in order to improve industrial process through resilient and flexible manufacturing systems. Process control and decision making in real-time are just two of the fundamental challenges that companies must face in order to keep pace with an increasingly competitive market. For this reason, predictive maintenance becomes the core of an industrial process, able to constantly monitor the system and make decisions before a failure occurs in order to improve the production, minimizing unplanned downtime and ensuring product quality.

The aim of the treatment is to design an IoT platform for predictive maintenance purpose, in which the acquired data are computed using the edge computing technology. Thanks to this technology, the services of the platform are able to elaborate data directly on the field, near to the plant where they are generated, reducing the amount of data to be sent to the cloud and solving latency problems. The platform has been designed to manage data coming from any machine able to collect data from the plant, thanks to the provision of sensors. However, its adjustment to resistance spot welding machine has been chosen as case study. At first, an overview of IoT technology is presented. Consequently, different computing data platforms are discussed and a solution for a predictive maintenance application in manufacturing production is described in details. In the last sections, the theoretical bases of the resistance spot welding process are analyzed and the related laboratory experience, aimed to collect the data directly from the machine, is explained. At the end of the paper the conclusions are drawn, in which future work will be proposed.

Introduction

The growth of greater market competitiveness has led industries to research and exploit cutting-edge technologies in order to increase productivity and product quality, while reducing costs. The achievement of these goals has been made possible by the focus on predictive maintenance regarding the production process, which can reduce unplanned down time, reducing production costs without compromising performance. Despite the increased complexity of the systems involved, the development of new maintenance strategies has also been able to increase the flexibility and reliability of the overall manufacturing systems.

This has been made possible by the parallel growth of technological innovations based on the use of artificial intelligence. Intelligence maintenance systems (IMS) and Intelligent prognostic and health management tools are just some of the examples belonging to this field, having respectively the task to provide decision support tools to optimise maintenance operations and identify effective, reliable, and cost-saving maintenance strategies to ensure consistent production. [1]

This exponential technological growth since the 1990s has led to the birth of the Internet of Things (IoT), a reality in which devices are able to communicate with each other via an internet connection and a cloud server. IoT is in fact a network of devices, where a wide range of devices with different functionalities and designs can exchange information with each other, process data and make decisions at run-time. This technology has a place in many fields, including manufacturing, home, healthcare and military. [2]

This topic will be discussed in more detail in the next section, as it underpins the issues discussed in this paper.

Chapter 1: IoT in Enterprise Organization

1.1 Internet of Things

The term "Internet of Things" was first coined in 1999, when Kevin Ashton defined a new method of device-to-device communication. [3]

Many definitions were given to the term IoT in the following years. Among them is the following, taken from the 2012 report by the International Telecommunication Union (ITU): the "Internet of Things" can be seen as "*a dynamic global network infrastructure, as such it can identify, control, and monitor each object on earth via the internet according to a specific agreement protocol, and through the interconnection of physical and virtual things based on the interoperability of information and communication technologies*". [4]

The purpose of this communication network is to enable different devices to exchange data and information at run-time in an autonomous manner. A representation of this concept is depicted in the figure below. The IoT device can be identified as any sensor with computational intelligence. Located in a system with an internet connection, it is able to communicate with any other device at any instant of time, as long as it also belongs to the same network. The location and transmission of data between the many devices belonging to the network is then allowed thanks to a solid IoT infrastructure, able to collect all the data exchanged and store them within the system, for future computation for the most diverse purposes. There are different types of networks, including Wi-Fi, Bluetooth, etc., and only devices equipped with a physical IP or with integrated wireless sensors are able to access them. They are called "smart devices" because of this peculiarity. [3]

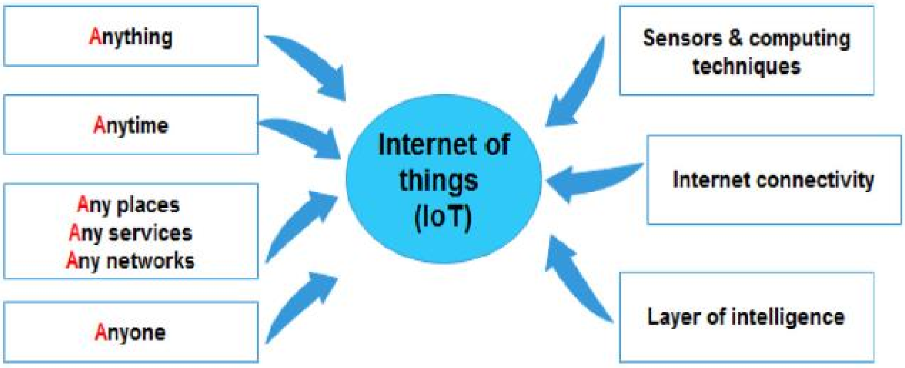


Figure 1 - Depicted example of IoT concept [3]

An example of an IoT system is illustrated in figure 2. The simpler case is presented: a smart home system. Three layers can be identified: the first one, is the physical one, in which all the smart device belonging to the system are included. The term "*device*" refers to all those sensors whose task is to translate the parameters acquired directly by the system into digital data. They take this name because they are integrated into the devices that host them. They act not only as inputs for IoT applications, but also as outputs, as they act as actuators. Their task is to translate digital data from the cloud into physical signals in order to trigger an action on the system they belong to. Devices belonging to this category can be recognised by looking at the figure below. They include smart refrigerators, microwaves, alarm systems, etc.

The second layer is composed by "*Gateways"*. They are edge devices with dual communication technology: having two types of channels, they are able to communicate with the upstream layer, the cloud, with one of them and exchange information with downstream devices, the physical system, with the other one. The communication with the upstream layer is allowed by integrated communication modules (for example Wi-Fi chips), hosted into such devices. These modules enable the communication of the devices to network gateways. An example of it is the Wi-Fi router depicted in figure 2.

Another functionality of the gateway devices is related to the computation of data. In fact, they are able to perform data operation as deduplication, data segregation, aggregation, clean up and edge computing. About this last topic, a more detailed description will be presented in the following chapters.

The last layer includes the "*Cloud*". It collects all the data coming from the plant and takes memory of them. Moreover, it hosts the IoT platform, the orchestrator of the whole system. Its aim is to deals big volume of data at high speed. It also hosts algorithms and computational features for deeper analysis of data. [5]

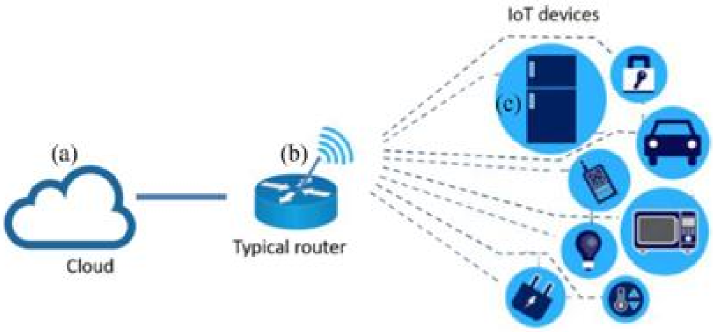


Figure 2 - IoT solution for smart home system [2]

Thanks to the enormous advantages of this technology and the possibility of applying it in numerous fields, it has also been introduced for the management of industrial processes in the manufacturing sector. The IoT has made it possible to create smart environments that can now independently monitor and control the production processes they host, improving performance and reducing costs. The monitoring network consists of several layers, similar to the previous example. The heart of the entire system is also in this case a massive platform, equipped with a data centre and a system of support services for computing the data collected. As previously stated, everything is made possible thanks to the use of numerous interconnected devices and integrated sensors able to connect to the network.

The possibility of a large number of devices to communicate, compute, control and sense the surrounding environment, leads to a huge problem: the handling of Big Data. Due to the implementation of this technology in a such enormous environment brings the generation of a massive volume of data, that requires high-velocity and high-computational power. Indeed, the term "big data" refers to "*large scale of data that demands new architectures and technologies for data management (capturing and processing) to enable the extraction of value for enhanced insight and decision making*". [6]

For this reason, being able to store and compute such a vast amount of data is still a challenge. However, in recent years great progress has been made in this direction. Nowadays, different and more performance tools are available in IoT application in order to extract information in uniform manner from heterogeneous data and to analyze them with reduced computing power and higher velocity. The advent of web 2.0 has prompted the development of new methods of communication, making it possible to integrate different types of data-gathering source, such as social media and IoT enable sensors, with innovative data analysis tools, such as Kafka and Spark, in a real-time processing, required to satisfy smart-environment applications' purposes. [3]

Chapter 2: IoT platform for IoT applications

In order to understand what is meant by an IoT platform it is necessary to first consider an IoT system as a whole. An IoT system is composed by a set of connected entities working together, including device belonging to the physical world such as sensors and actuators and to the virtual one, such as software services, enabled technologies and algorithms. The purpose of the system is unique and is based on the computation by the virtual instruments of the data collected from the physical devices in real-time. Therefore, the virtual and physical worlds operate together in an orchestral way in order to achieve a common goal. [7]

Many aspects that make up an entire IoT system have already been described in the previous chapter, and the image below is intended to clarify this concept.

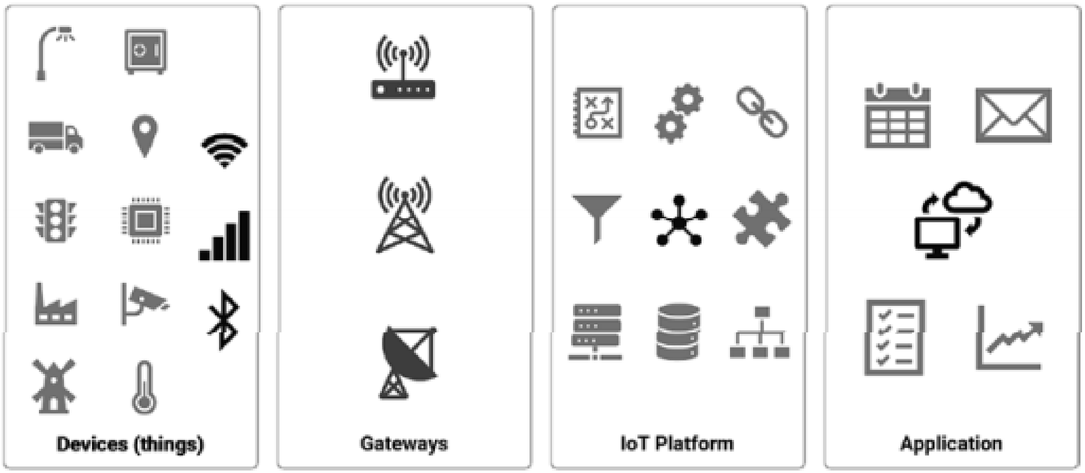


Figure 3 - Representation of an entire IoT system [5]

The discussion in the first section of this chapter will focus on the detailed description of all the main building blocks of an IoT platform, leaving to the following section the analysis of how the platform integrates with the other layers of the system and works in collaboration with them. A first representation of what the structure of an IoT system looks like is depicted in figure 4: the cloud layer is the layer of first interest, as it is the one where the IoT platform resides.

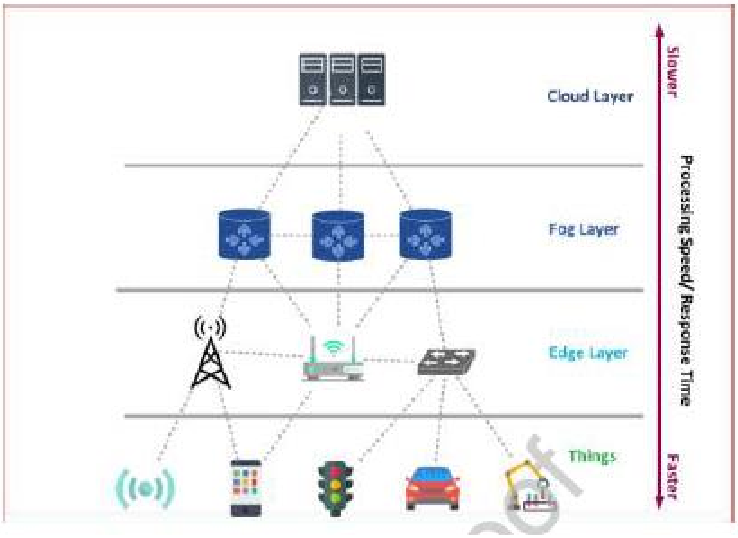


Figure 4 – A general architecture of an IoT system [8]

2.1 A general IoT architecture

As mentioned above, the structure of an IoT platform is analysed below, shown in its entirety in the image below.

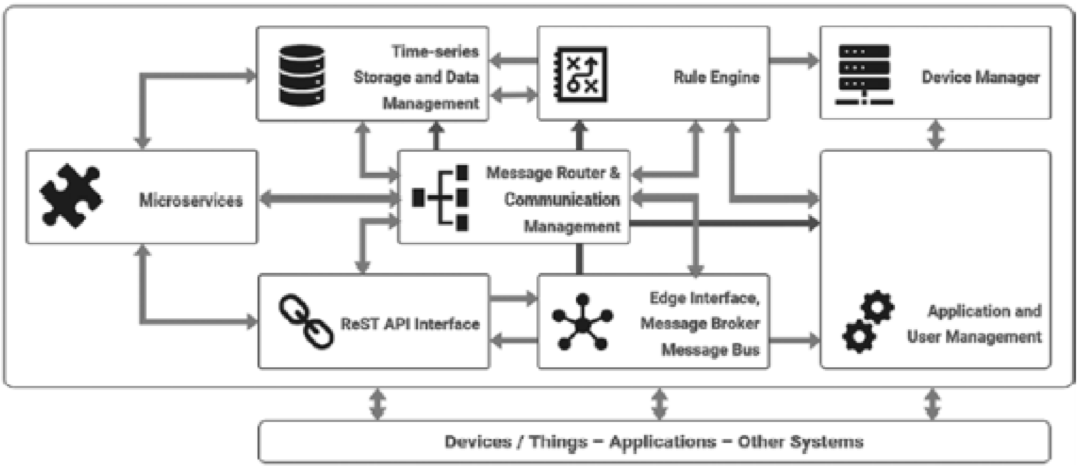

Time-series Storage and Data Management
Rule Engine
Device Manager
Microservices
Message Router & Communication Management
ReST API Interface
Edge Interface, Message Broker Message Bus
Application and User Management
Devices / Things – Applications – Other Systems

Figure 5 – A general IoT platform and his main blocks [5]

A general IoT platform is composed by different components, each with a specific function. The main components are:

- Edge Interface, Message Broker and Message Bus
- Message Router and Communication Management
- Time-Series Storage and Data Management

- Rule Engine
- The REST API Interface
- Microservices
- Device Manager
- Application and User Management

Most of them are not essential for proper functioning of the platform, but each of them is useful for a scalability, flexibility and reliability platform. The core functional components are the device interface and message broker, the message router and communications module, the data storage, the rule engine and the device management. Thus, it will be analysed them first**.**

Edge Interface, Message Broker and Message Bus: it is responsible of the coordination of messages in the platform. His role is essential since it is able to communicate with the physical system, receiving from it the outputs coming from sensors and devices installed on the plant's machines, and putting the data collected in such way into the common message bus, readable by the other components in the cloud. Hence, it is the bridge from the real world (the physical system) and the virtual one (the cloud hosting the platform). Since devices and sensors are different from each other, they also communicate with different protocols and through different technologies (e.g. Wi-Fi, Bluetooth and so on). So, the Message Broker has to be able to communicate over a multiple module and uniform the coming data in a unified manner.

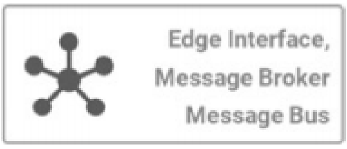


Figure 6 – Message broker block [5]

Message Router and Communication Management: strictly connected to the Message Broker, his task is to take the data from the main message bus, re-elaborate and rebroadcasting them into the message bus. In particular, it performs action as enrichment data with more context, refinement, publishment of additional information useful for other components to read and understand messages, conversion of data format and deduplication of messages (ability to eliminate duplicate messages from redundant data since useful).

Figure 7 - Message Router and Communication Management block [5]

Time-Series Storage and Data Management: useful for storing the data sensed from the system, now accessible on the message bus. The data present on the message bus are in sequential order, so the job of the Data Repository is to store data in suitable order for the platform's components. The type of the acquired data is in time-series one, hence it will be necessary a pre-procession action to convert them in a useful manner.

Figure 8 - Time-Series Storage and Data Management block [5]

Rule engine: the executive core of the whole platform. It acts following pre-defined rule and monitors the message bus constantly. Reading messages and events across the virtual system, it is able to take action and execute algorithms. Every time the rule engine triggers, an action is performed in the platform. As well as acting according to the impulses generated by the physical system, the rule engine performs the generation of data from the cloud to the plant, sending them as inputs to the actuators. An example can happen when a component of the plant breaks down due to an unforeseeable event and a light on the plant is turned on in order to alert the operator. In this case a message warning of rupture is put in message bus. When the rule engine reads the message, it triggers and performs an answer (the command to turn on the light). The information coming from the engine is put in turn in the message bus, available to be read and sent to the actuators. Consequently, the role of the rule engine is an example of the biunivocal communication between the plant and the cloud.

Figure 9 – Rule engine block [5]

Device Manager: this component is useful for manage devices and control the status of them. It provides a list of devices, with related functionalities such as network conditions, their status, access key, battery level, details, and so on.

Figure 10 – Device Manager block [5]

The REST API interface: useful to put in communication programs, devices or utilities that do not need to communicate constantly or in real-time with the platform. An example of such device is the temperature sensor, that sends temperatures values each time range of 15 minutes. No being in connection in real-time, the communication protocol requested to send data to the platform is a simple HTTP. As a consequence, this communication over REST API interface imposes coordination between the REST API's work and the message broker's one. The REST API is also able to access the time-series data repository to give back a response to the sensor.

Another powerful functionality of this component concerns smart device in autonomous system. In this case, the REST API works in collaboration with the rule engine: when a specific event happens, the API interface allow the involved application to trigger the rule engine, in order to execute a pre-defined operation. An example of it, is the smart lock. When a sensor detects suspicious movement at the door, the application triggers the rule engine for the execution of predefined workflow (notifying security, sending message to the user, enabling acoustic alarm, etc.).

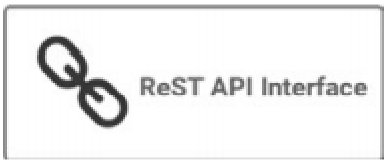


Figure 11 – REST API Interface bock [5]

Services and Microservices: Services are the specific applications that differentiate the platform from the others. They can be seen as the identity of the platform since explain the purpose for which the platform is created. Microservices are instead auxiliary services such as payment services integration, email notification, verification and so on.

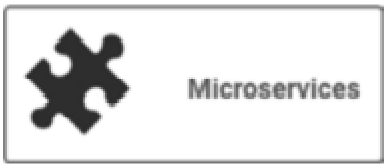


Figure 12 - Services and Microservices block [5]

Application and User Management: It is related to application and user. Indeed, provides functionalities such as access keys, credentials with utilities and passwords, login and so on. Its use is therefore similar to that of the device manager but referred to users and upstream applications**.**

Figure 13 - Application and User Management block [5]

2.2 Data computation

A challenge in the realization of a platform for IoT applications is the computation of acquired data. Indeed, the communication's efficiency is strictly influenced by the data processing methodology, resulting in the optimisation of the decision-making process. [9]

Hence, before explaining the working of the platform and how the various components are connected and work in coordination with each other, it is necessary to make a premise by introducing the concepts of "*Cloud Computing*", "*Fog Computing*" and "*Edge Computing*".

The term "Computing" refers to the management and processing of information. There are three broad categories, named according to the layer in which this process takes place, mentioned above. The Cloud Computing is a data processing methodology operated by the cloud, resulting in huge power computation and centralized data storage, shared by multiple users. Fog Computing is the methodology in which the computation is performed by the fog layer, an interface between the cloud and the smart devices located in the edge layer, composed by fog nodes with power computation. The last of them is the Edge Computing, in which the processing is done in the edge layer, by the sensors and devices themselves or in the gateway devices near to them. The concept behind this strategy is to operate as close as possible to the source of the data, without sending it to a third-party system. As a result, the architecture is no longer centralised as in Cloud Computing, where all information is brought together in the cloud and all operations are carried out there, but takes on the characteristics of a decentralised system, as already anticipated in the architecture of Fog Computing. The computational performance of the entire platform is thus improved, as data and its processing can take place instantaneously, since the waiting time occupied by sending data has been removed. [9]

In order to better understand why the choice has fallen on edge computing technology, the current section aims to present an overview of the differences between the existing data processing technologies, by presenting the main features of each of them.

2.2.1 Cloud computing

Cloud computing is an on-demand network access technology based, able to share data and resources to compute by servers, services and applications.

Platform based on cloud computing technology is profitable for business owners, thanks to his overall lower cost. Indeed, it does not have an up-front cost and the operating cost is reduced, since all the data and the computing resources can be allocated and removed on the cloud on demand. The maintenance costs are also minimized. Another peculiarity is the high scalability of the system, since it is easily expandable. Moreover, multiple devices can easily access services in the cloud through internet connection. In order to understand the capabilities of this technology, a comparison about managing the computation of data stored in datacentre between a Cloud Computing approach with a traditional one will be done.

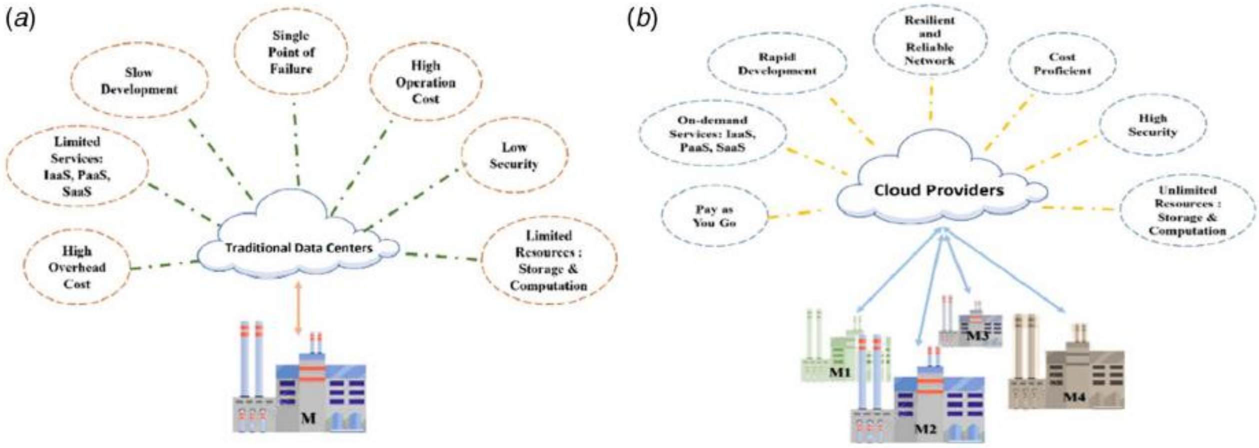


Figure 14- Comparison between cloud computing approach with a traditional one [1]

The peculiarity of cloud manufacturing is in englobe all the system, from the process of the product, simulation, design of the system, and so on. into a service on a cloud, accessible by users. Thanks to the service platform, users are able to request services at any stage of the system, monitoring and analysing all the aspect of the plant, in order to achieve a common goal. [1]

The desire to exploit a technology such as the one presented above stems from the need for companies to organise and manage production processes that handle an enormous amount of heterogeneous data. It is therefore important that the data generated by such a complex and articulated system is always accessible in its entirety, at all times, since the computation of just one part of it often leads to a wrong decision, responsible for a drop in the performance of the entire production system. In order to manage continuously generated data of such a vast nature, the Information and Communication Technologies (ICT) proposes as a traditional solution, the installation of an end-system, in the terminal areas of processes, capable of collecting data, storing them and analysing them. However, installing an on-site solution requires a massive effort in terms of power, security and capital investment. Since the construction industry is made up of the vast majority of small and medium-sized enterprises, a large investment as the sole prerequisite for such an innovative service is not sustainable for most of them, considering that it would have to be applied to every single process that makes up the production system on which the company is based. The advantage of cloud computing is that there is no need to install a dedicated on-site system to compute the data collected, since the aim is to create a re-usable infrastructure in order to reduce initial and maintenance costs. [10]

All the aspect involved in the Cloud Computing technology are summarized in figure 15, which underline its five distinctive features, the types of service models that the cloud provides to the user and the four different deployment models.

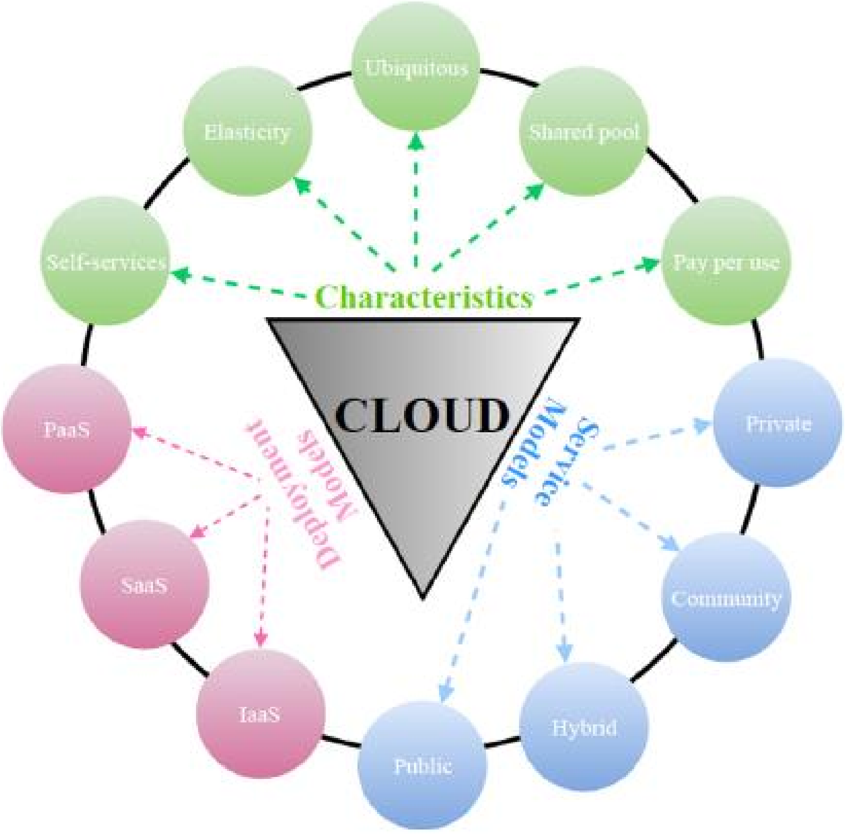


Figure 15 - Overview of cloud computing technology [10]

The five characteristics showed in green are already mentioned previously:

- Ubiquitous: users can access to the services of the cloud anytime and anywhere. The multiple locations nature of the computing infrastructure allow connection from any device capable of reaching the network.
- Shared pool: users can share the same infrastructure using different applications, but the security and the privacy of all of them is always guaranteed.
- Elasticity: the cloud is always able to satisfy an increase or decrease of computing request by multiple users.

- Self-service: the cloud is made available by the user who wants to use it, automatically, guaranteeing a fast service and instant deployment. Cloud providers allow the sale of the cloud through websites without their intervention, also providing a web interface, so that access to the cloud is immediate for those who want to use the service.
- Pay per use: by exploiting the pay-as-you-go feature, users are allowed to pay only for the services they benefit from, although the cloud provides a very wide variety of services.

The architecture of a cloud computing platform is composed by three service models (or layers). "*Infrastructure-as-a-Service*" is the first one, also called the "provider domain" or technology layer, divided in two sublayers: resource and virtualization. In the resource layer, are included all the resources of the plant that can be offered to users as a service. The aim of the virtualization service is instead to virtualize the resources and package them into services. Moreover, thanks to it, users pay to rent a virtual storage in order to allocate the resources. "*Software-as-a-Service*" is the core of the platform, the so called "enterprise domain", in which the global service layer is a centralize virtual system, that provides dynamic cloud computation services, including monitoring, quality controlling, user management and so on. Software developed in order to achieve these purposes are available on a pay as you go basis, reducing the cost of the cloud. "*Platform-as-a-Service*" is the "application layer", useful to provide a platform in order to build applications. Then, computers and clients are enabled to access this layer, through which they are put into communication with the manufacturing cloud services. Then, it is the interface of the platform for the users. [1] [10]

Depending on which category is allowed to access cloud resources and the methodology used, cloud services are classified into four categories according to the type of deployment: a public, a private, a community or a hybrid service. The first one is suitable for small businesses, where a public cloud can be reached via Internet by many users, who have access to the data centre where the applications reside. The control and the entire management of the system is entrusted to the cloud provider. The private service is useful to keep track of sensitive information, accessible for this reason by one or few organizations via internet. For example, government institutions operate with private service in order to handle their large businesses with sensitive data belonging to national infrastructures. Less exclusive than the previous one is the community service, which allows access to a community of companies that need to share the same cloud infrastructure, working together to achieve a single goal. The last deployment is the hybrid cloud, a combination of one or several types of deployment. This solution is suitable for platform that handle different types of information, some sensitive and some non-sensitive. In this case, the sensitive ones are accessible only by a limited number of users, while the others are accessible to all users. [10]

The centralised format in which the platform operates and the accessible huge datacentre brings numerous advantages to users, such as economic benefits and high manageability. However, the disadvantages are also considerable, should it be necessary to implement a technology that can compute in run time data from devices allocated in locations far from the cloud. The latency introduced in the communication between users and the server would be too high to meet the need for an instant response due to long distances that the sent messages have to cover across the wide area network (WAN). Moreover, problems of high energy consumption during condition of operation and the need for a large amount of capital to be invested in the implementation of the platform occurs when dealing with such technology. [1]

A geographically distribution of small datacentres is a possible solution for these challenges and leads to new technology: the so called "*Spanning Cloud Computing*" (SCC). Having a set of datacentres leads to a reduction of jitter and latency of signals and a greater advantage exists if one of them is no longer able to function. Unlike the traditional cloud, if an external attack were to bring down a datacentre, the entire system would not stop working, but users would still be able to use the services of the platform thanks to the presence of the other datacentres. The resulting effect is an improvement in cloud performance and the introduction of an additional layer of security. However, having more data centres puts the system at greater vulnerability, as there are more access points at which hackers can attack. In addition, having special software and tools to control and coordinate a more complex platform entails higher management costs, making this technology more difficult to maintain. Therefore, this solution is a double-edged sword, as it induces an improvement in performance but at the expense of increased costs. Although some aspects have been improved, the platform still does not fully meet the requirements of IoT applications such as tight latency and low jitter. [10]

Further studies in this direction lead to the "*Cloud-based Content Delivery Network*" (CCDN), a methodology that brings servers closer to the user, thanks to the use of edge servers allocated in “Point-of-Presence” (PoP), that are servers allocated at strategic points in the system. The resulting reduction in traffic stems from the cloud's ability to find the best possible path to send information to the most suitable edge server to receive it, thus reducing jitter and response time. Despite these improvements, the minimum requirements are still not met for use in the IoT environments which have to handle a huge amount of data in real time. [11]

2.2.2 Fog computing

The main problems encountered in cloud computing are largely solved by the development of a new methodology called “*Fog Computing*”. A definition of Fog computing was provided in 2012 as “*an extremely virtualized environment that delivers networking, storage, and compute resources between outdated CC information centers, usually, but not entirely situated at the network edge*”. [12]

Then, fog computing can be considered an extension of Cloud Computing, able to bring the cloud closer to the network edge. Indeed, the Fog Computing architecture allows the computation of generated data by devices close to the edge layer through fog nodes, where information is locally stored and analysed, without having to be forwarded to the cloud server. Being able to process and store data in a layer between the cloud server and end devices, the resulting effect is the ability of the system to provide answers more quickly, leaving the possibility to freely access the data centre in the cloud when needed. In this perspective, Fog Computing provides a single service model, called “*Fog-as-a-Service*”, where service providers allocate fog nodes geographically distributed across the platform, which the aim to collect data, process them and improve the performance of the system. This feature makes the approach fully distributed, unlike Cloud Computing which is fully centralised. Every node therefore has similar computational capabilities to the cloud, but far less powerful: it is still the cloud server's job to perform more complex tasks and make more complex decisions. Fog nodes are smart devices such as switches, routers, smartphones, base stations, network device management and so on. Figure 16 shows a Fog Computing architecture, composed by three layers: the cloud, as in Cloud Computing, the fog nodes layer, where fog node servers are present and are able to communicate each other and the end device layer, where smart device are present, able to acquire data and sent them to the upper layers. [13]

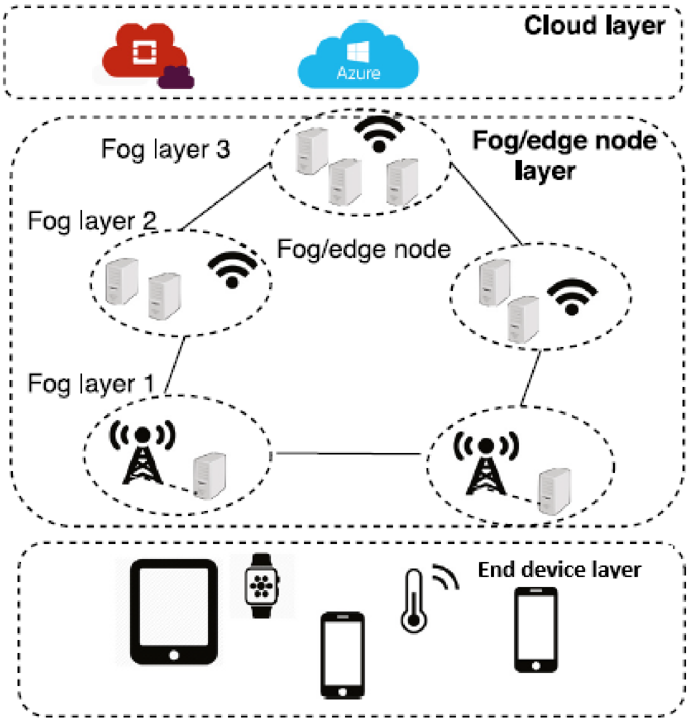


Figure 16- Example of fog computing architecture [11]

Looking to the “Fog/Edge node layer”, it is clear that there is a vertical structure between the devices belonging to it, depending on how they are linked together. Fog node is a term that can encompass several different devices: a gateway, fog device or fog server. Fog server is the powerful of them, since it manages the others and controls them. Fog devices are instead linked to a cluster of sensors belonging to the system, both virtual and physical, to which the device refers. Only fog devices belonging to the same cluster and linked to the same fog server, are able to communicated each other. However, it is often the case that a single application requires the use of several groups of fog devices, so although they do not communicate directly with each other, they are able to exchange information via the servers to which they refer. [13]

Thanks to this technology based on fog nodes, the client can have real-time response from the system even in applications very sensitive to latency. [13]

Other advantages of a structure like the one presented above are:

- Adaptability: the network ca be easily extended and the disseminated fog nodes can operate along every network equipped with dedicated sensors
- Real-time communications: the analysed architecture is specifically designed to enable simultaneous communication

- Physical distribution: being a decentralized architecture, services and applications can be hosted in any fog nodes without limitations
- Compatibility: thanks to the adaptability pf fog nodes, the communication with other platforms and networks can be achieved in simple manner
- Heterogeneity: the possibility to configure fog nodes to adapt the system to numerous device and platforms, leads to a platform hosting very heterogeneous devices and services
- Provision for flexibility: as a consequence of the feature before. The platform has also the ability to connect mobiles, thus providing flexibility techniques. [13]

Thanks to his main features, the Fog Computing finds several applications in very different environments. Some of them, are depicted in the figure below:

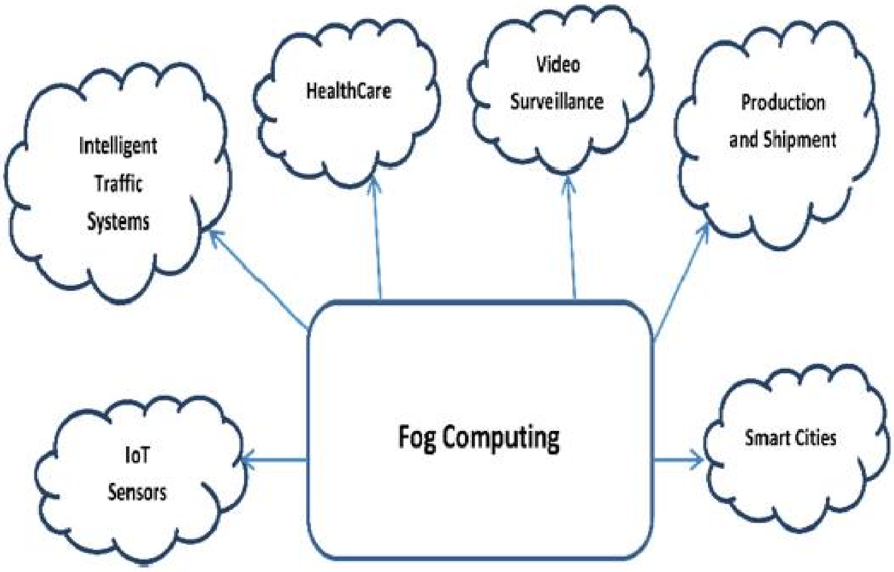


Figure 17 – IoT applications exploiting Fog Computing [13]

2.2.3 Edge computing

Following the same line of thinking that led to the development of Fog Computing, a new technology has been proposed, called "*Edge Computing*". The basic idea is the same: to bring the computation of collected data as close as possible to the end of the system through the use of devices with processing capabilities, thus lightening the load of information that must transit through the cloud. The devices mentioned fall into three categories: end devices, including smart objects and mobile phones; edge devices, which include bridges, switch boxes and hotspots, and edge servers, those who have the highest processing capacity. [13]

Edge Computing is widely used in IoT applications, as there is a need to have a response from the server in the shortest possible time, a requirement that cannot be met by a cloud-based server, which, having a centralised architecture, is usually at a fixed location, away from data sources. By being able to filter, analyse and process data close to where it is generated through Edge Computing, the system is able to make crucial decisions in real time, while achieving low latency and performance requirements. However, this technology does not give up the cloud presence, as although the devices (single board computers, smartphones, routers, switches and so on) are able to collect, pre-process and perform an initial analysis of the data, they are not able to extrapolate more complex results from a more in-depth study, which is made possible through the use of approaches based on optimisation and machine learning. [11]

The advantages that a decentralized architecture offers, is reduction on jitter, latency and load on the network, increasing security thanks to store data in on-site infrastructure. However, researchers have very divergent views on the definition of an edge and its location, as it strongly depends on the context in which it operates and the application for which it is required. Depending on how consumers and data providers varies, the edge assumes a different location, as the logical border of the network changes. For example, in the context of autonomous driving, the vehicle, the physical device, is the edge of the network, where the place where the data is collected and processed is the same, whereas in a CCDN architecture, the server is the edge of the system, as the task of processing the data is entrusted to it. [11] In both cases, the data computation takes place in the edge of a system but does not necessarily imply that it is Edge Computing. Consider, for example, the case where a small server located at the edge of the network performs the function of data processing within a Cloud Computing architecture, such as in the cases of SCC and CCDN, or consider a mobile phone on which an application of Cloud Computing runs. In both cases we are not in an Edge Computing architecture, although the object belonging to the edge layer, performs the function of computation, being between the cloud and the IoT sensor or the application. The previously mentioned case of an autonomous vehicle is different, where the device, belonging to the edge of the network, is certainly able to act depending on the result produced by the analysis of the data, but it is also the sensor itself that produces and stores them. Therefore, the discriminating factor in edge computing is that data computation, analysis and storage must take place as close as possible to the place where they are acquired, without precluding the possibility of sending this data to a remote server that can process it in more detail for more complex applications. [13]

In one respect, Fog and Edge computing can have many similarities and are often mistakenly used as synonyms, as in both cases the information collected is sent to devices close to where it is generated. Specifically, Fog Computing is a distributed infrastructure in which certain processes or application services are managed at the edge of the network by an intelligent device, while others are still managed in the cloud. It is an intermediate layer between the cloud and the devices in order to enable more efficient processing, reducing the amount of data that needs to be sent to the cloud. Fog Computing requires the use of an external node or gateway, while Edge Computing processes data directly in the devices themselves. Instead of using a gateway, several smart devices can handle the data processing. [14]

Fog computing has the advantage of allowing a single, powerful device to process data received from multiple endpoints and send the information exactly where it is needed. Compared to the Edge, Fog Computing is more scalable as it allows a centralised system to take a broader view of the network by having multiple data points feeding information. On the other hand, Edge Computing pushes the intelligence, processing power and communication capabilities of an edge gateway or appliance directly into devices such as programmable automation controllers, with ultra-low latency for run-time applications. Linking sensors to programmable automation controllers that manage processing, communication and other tasks, each device in the network will play its own role in processing information, avoiding of perform most of the work in a centralised server. [14]

No one technology is better than another, it depends on the type of application you need. The biggest advantage to be gained from cloud computing over the other two is that it is more suitable for long-term in-depth analysis of data, but it has to be replaced when a rapid analysis is required for real-time response. Moreover, Cloud Computing is the least secure from a security point of view, since it is the only technology that requires continuous internet access: the other two can work even without the Internet, reducing as a consequence the time period in which they can be subject to hacker attacks. Another advantage of Cloud Computing is that all data are on the same server and therefore easy to handle. On the other hand, if something goes wrong, it brings down the entire system, while in Fog Computing, since data is distributed among nodes, downtime is minimal: if a node crashes, other nodes remain operational, making it the right choice for use cases that require zero downtime. The same result is achieved in Edge Computing. [15]

Chapter 3: Design of the IoT platform for predictive maintenance purpose

3.1 The evolution of maintenance paradigm

In order to increase the productivity of a production process, minimising unexpected waiting times, and maintain the quality of a product without raising production and operating costs to remain competitive in an ever-growing market, industrial companies resort to the use of maintenance operations aimed at predicting unpleasant events that lead to delays in production and a drop in earnings. These maintenance operations are designed in order to provide support to the systems that deal with them, aimed at make run-time decisions to optimise production. They are called Intelligent maintenance systems (IMS) and for many years have been the subject of much research aimed at finding the right strategies to achieve a balance between increasing productivity while maintaining reliability. Their role becomes substantial especially in systems composed of numerous machines, as the more devices there are, the greater the risk that one of them will break down, causing a serious economic loss to the company. Maintenance operations therefore have the task of improving the monitoring of machines in order to control their operational status. Over the years and with the advent of Industry 4.0, the field of maintenance has evolved rapidly. Looking at the figure below, it is possible to divide this process into four periods, expiring with the parallel evolution of the entire manufacturing field. [1]

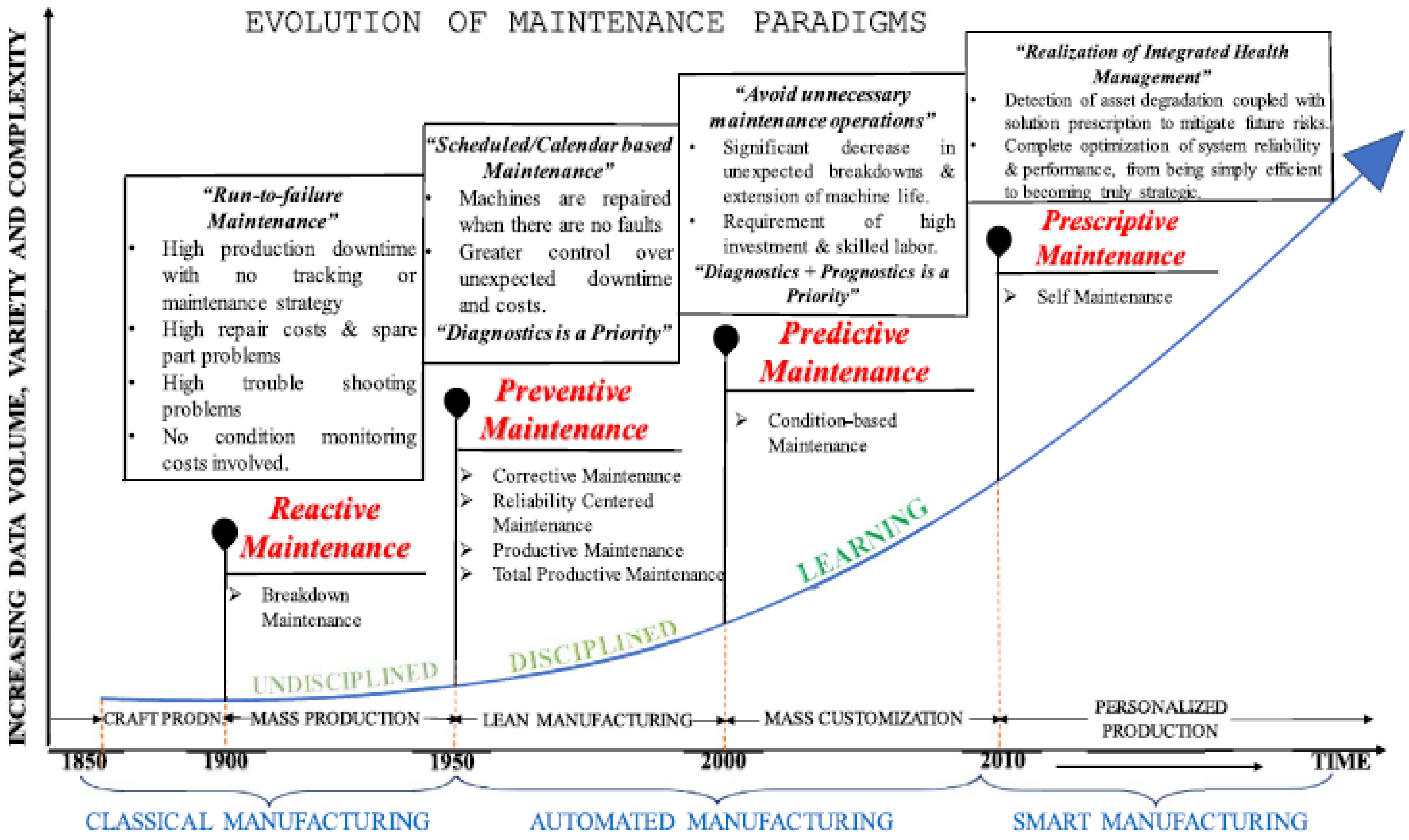

EVOLUTION OF MAINTENANCE PARADIGMS
INCREASING DATA VOLUME, VARIETY AND COMPLEXITY
"Run-to-failure Maintenance"
• High production downtime with no tracking or maintenance strategy
• High repair costs & spare part problems
• High trouble shooting problems
• No condition monitoring costs involved.
"Scheduled/Calendar based Maintenance"
• Machines are repaired when there are no faults
• Greater control over unexpected downtime and costs.
"Diagnostics is a Priority"
"Avoid unnecessary maintenance operations"
• Significant decrease in unexpected breakdowns & extension of machine life.
• Requirement of high investment & skilled labor.
"Diagnostics + Prognostics is a Priority"
"Realization of Integrated Health Management"
• Detection of asset degradation coupled with solution prescription to mitigate future risks.
• Complete optimization of system reliability & performance, from being simply efficient to becoming truly strategic.
Reactive Maintenance
➢ Breakdown Maintenance
Preventive Maintenance
➢ Corrective Maintenance
➢ Reliability Centered Maintenance
➢ Productive Maintenance
➢ Total Productive Maintenance
Predictive Maintenance
➢ Condition-based Maintenance
Prescriptive Maintenance
➢ Self Maintenance
UNDISCIPLINED
DISCIPLINED
LEARNING
CRAFT PRODN.
MASS PRODUCTION
LEAN MANUFACTURING
MASS CUSTOMIZATION
PERSONALIZED PRODUCTION
1850
1900
1950
2000
2010
TIME
CLASSICAL MANUFACTURING
AUTOMATED MANUFACTURING
SMART MANUFACTURING

Figure 18 – Evolution of maintenance paradigm [1]

The first phase concerning maintenance operations arose as a result of the second industrial revolution, in the second half of the nineteenth century. It is a classical industry, focused solely on mass production. Since the only aspect that is analysed is the production system itself and producing as much as possible, the focus has not yet shifted to how to prevent unexpected downtimes in a production system. Maintenance operations therefore only take place when a breakdown occurs, resulting in high repair costs, but no capital expenditure on monitoring the machinery. For this characteristic of reaction to failure, it is referred to as "*reactive maintenance*". [1]

With the advent of automation, maintenance strategies began to emerge, aimed at carrying out repairs on machines before faileures occur. The "*preventive maintenance*" had therefore the task of operating preventively on machinery, performing monitoring and prevention operations such as inspection and evaluation on the most critical devices in the system. One example of a preventive strategy, is to perform repair actions periodically on the device on the basis of an evaluation parameter, e.g. the mean time between two breakage events. Another evaluation parameter on the basis of which maintenance actions can be performed could be the life cycle of the machinery considered: through experiments or consideration of the history of previously acquired data, it is possible to estimate the lifetime of a component. Based on this, it is then sufficient to choose a shorter time interval in which to operate, thus reducing unexpected downtime. [1]

In the modern age, maintenance techniques have evolved further with the use of systems to measure the operating conditions and status of machines during their lifetime. Because of this peculiarity, they are called "condition based maintenance" (CBM) techniques. This strand includes the "*predictive maintenance*", which aims to avoid operating when it is not strictly required, reducing system downtime as much as possible. Through a monitoring system designed to constantly assess the status of the machinery and specific tools capable of predicting the behaviour of the machine in the near future, it is possible to estimate the time interval in which the probability of failure is greatest. In this way, maintenance interventions are reduced to a minimum, saving costs and improving the consistency of the entire system. It is within this paradigm that the primary objective of the platform under investigation is placed. However, a further step has been taken in its design, as it is on the basis of the life-cycle assessment of the system components that the IoT platform makes decisions on what action to perform such as operation planning. This ability is one of the defining characteristics of modern maintenance techniques belonging to the so called "*prescriptive maintenance*" paradigm, the latest achievement in this field thanks to the arrival of smart manufacturing. Prescriptive maintenance paradigm is based not only on mere prediction of failure but also on the decision making such as operation planning depending on it and on the capability to find solution to completely avoid break down, bringing the downtime to zero. As the goal of this technology is very difficult to achieve, a lot of research is being carried out in this area, which is still growing. However, at present the intelligent maintenace systems that have emerged are able to provide users with powerful tools for diagnostics and decision support, to such a great extent that they are considered to be the basis from which to arrive at new prognostic strategies, that will be available to next-generations. This research area focuses on develop new tools in order to enable self-awarness, improve system monitoring, enabling "Prognostics and Health Management" (PHM) to all sectors and reducing as a consequence waste products, eliminating downtime. [1]

In figure 19 the main concepts of IMS are highlighted.

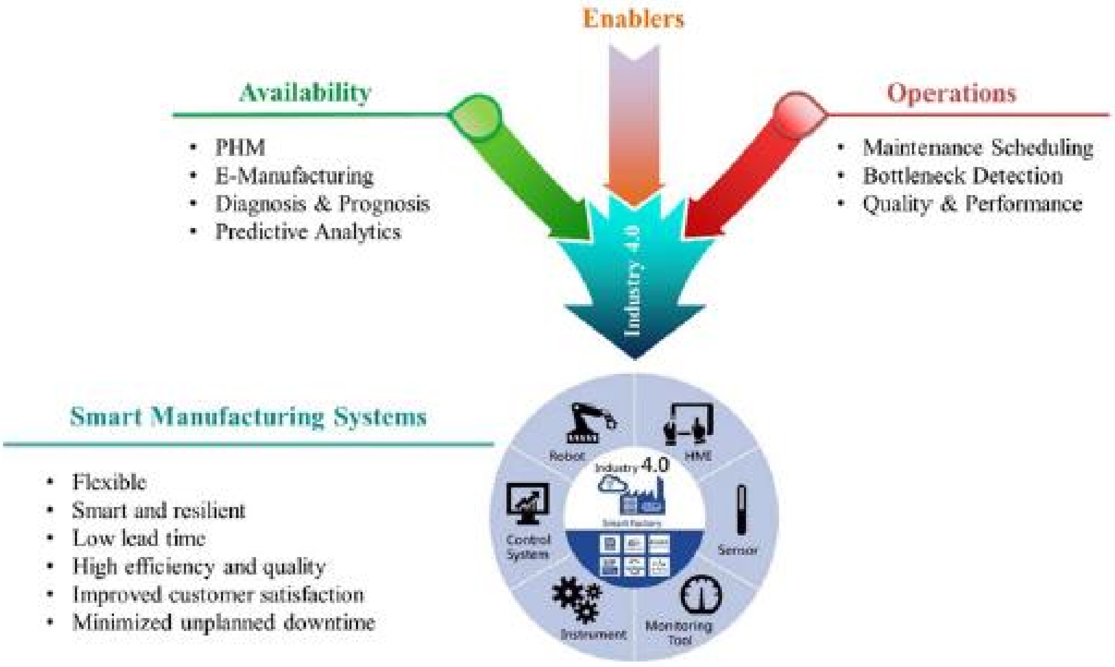


Figure 19 – Main concepts of IMS [1]

The pillars of a system belonging to such an advanced and innovative sector are the availability of components belonging to the systems, the optimization of operations and the enabling technologies. Thanks to PHM and powerful data processing tools, the availability of the systems can be maximaized. The dynamic modeling of the system is the main fature to optimized operations and an important role in minimising the negative impact on production of waiting time are played by maintenance events based on scheduling algorithms. All this is possible thanks to the developed tchnologies used as enablers of IMS, capable to put together conventional maintenance operations with innovations introduced by the digitalized smart manufactury. Through the predictive maintenance paradigm, it is not possible to avoid the degradation of the devices that make up a system, but it does have very powerful tools that can monitor their operation during use and estimate their wear and tear in the future. [1]

The most important of these is the aforementioned "Prognostics and Health Management" (PHM), having as primary objective to provide advance warning of component malfunctions and criticalities. It was introduced for the first time in 2006 and is defined as "*systematic approaches for maintenance and asset management, which utilizes signals, measurements, models, and algorithms to detect, assess, and track degradation, and to predict failure progression*". [16] [17] It is composed by six steps, each of them illustrated in figure 20:

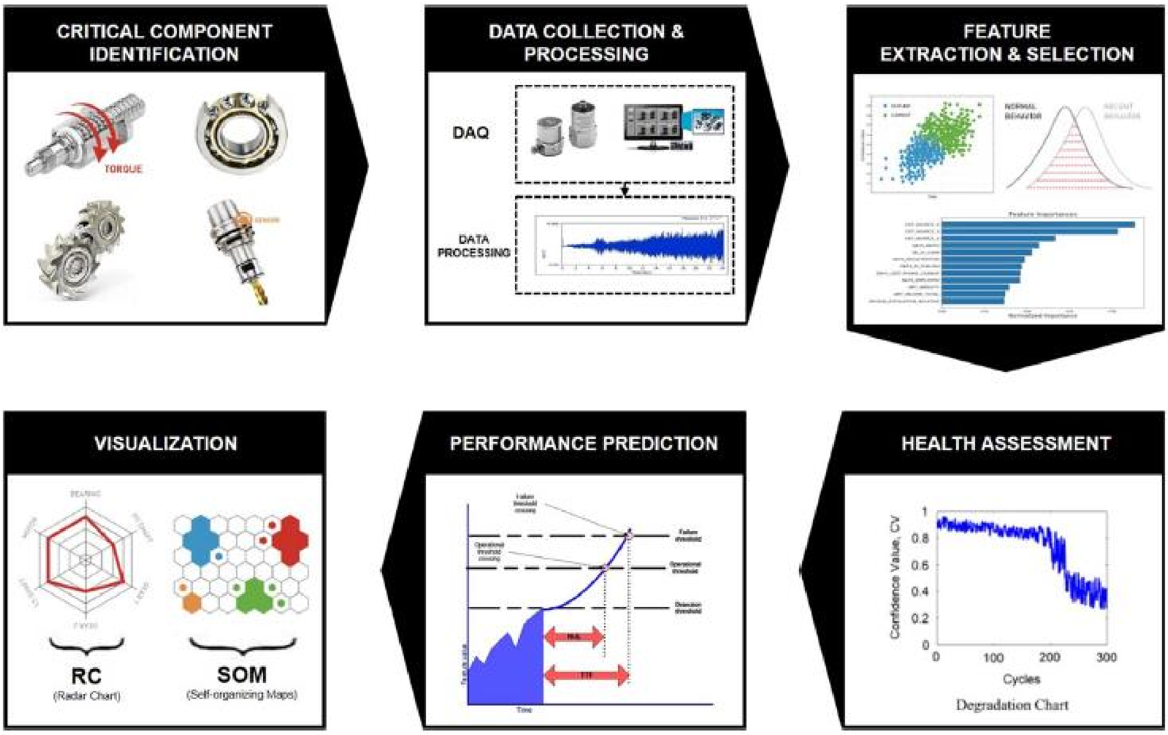


Figure 20 – Six steps composing PHM [1]

The procedure shown in the figure follows a sequential order indicated by the arrows. In this order: [1]

- Critical component identification: the first step is to identify components with a greater criticality or those that have a greater weight in the productivity and availability of the system. Once identified, monitoring and data analysis systems must be applied to them, as they are worthy of greater attention.
- Data collection and processing: in this phase, all possible data is acquired from the machines, sensors and controllers, so that, after a pre-processing operation such as filtering, outlier removing and data cleaning, it can be analysed in detail.
- Feature extraction and selection: Once acquired, data are converted from raw data into information that can be read and interpreted by dedicated algorithms. Not all the data taken is stored, but meaningful data is extracted from it in order not to burden the data flow and future computations. For this selection, special tools proposed by the researchers are also used.
- Health assessment: the objective of the fourth step is to establish the health of the machines under test, in order to understand the probability that the system may fail and due to which component. From the result of this analysis, a failure index is obtained, which expresses the probability with which a failure condition can occur.
- Performance prediction: depending on the index resulting from the previous step, a prediction of the machine's performance is made using prognostic algorithms.
- Visualization: is the last step, as it has the task of showing the operator or users the results of the analyses carried out so that it can support in decision-making.

3.2 Design of the platform and its functioning

After a general description of what an IoT platform is and from which component it is formed, it is possible to understand the logic and the functioning behind the design of an IoT platform for predictive maintenance. During the analysis of it, most of the components introduced in the previous section will be recognized immediately, since very similar to them, others will be little different, due to the necessity to obtain more detailed and complex performances.

The intention behind the design of this architecture is to create a platform capable of predicting the behaviour and trend of a physical process, through the computation of data acquired at run-time, while also taking into account all the historical data acquired during the phases preceding the current one. This latter feature ensures greater prediction reliability and a more precise and reliable estimate of future trends. Being multi-cycle industrial processes, it is possible to predict the future behaviour of the whole system, with the prediction of failures and alarms, by creating a data-driven model based on historical data. [18]

In order to achieve the goal of a platform with the capabilities to avoid unplanned downtime, to predict misbehaviour and wear of components, the edge-computing technology was chosen, as it is more powerful than the others for the reasons seen in the previous chapters. In the figure below, the design is shown in detail:

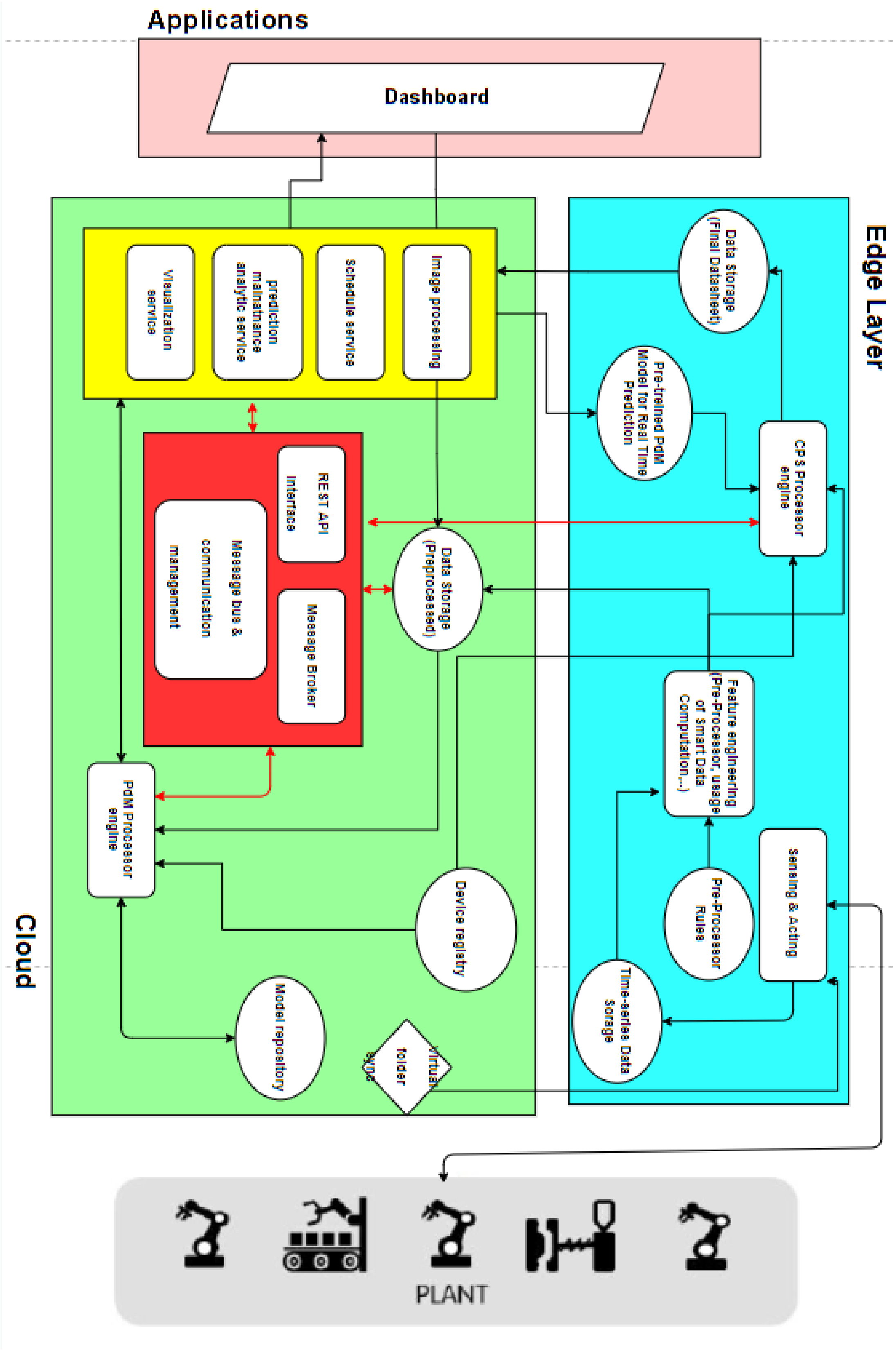


Figure 21 - Design of IoT platform for predictive maintenance paradigm

At first glance, the platform is divided into four sections, each in a different colour to make it easier for the reader to understand. However, it would be a mistake to think that this clear-cut graphic division reflects reality. In fact, many blocks belonging to different sections are constantly communicating with each other, exchanging information at run-time. This intricate network of communications is highlighted by arrows within the platform, many of which are bi-directional, indicating the mutual exchange of data between the components involved.

The subdivision into sections is therefore intended to divide the architecture of the platform into layers, in order to clarify the level to which the components belong. With image 3 in chapter 2 in mind, it is possible to identify a correspondence between the various layers.

The "*Plant*" is nothing more than the physical system from which the platform acquires data. This is why it is defined as the downsteam layer, and to this layer belong all the sensors, actuators and, more generally, all the machines involved in the process under examination, capable of producing data and information that can be analysed. It communicates directly with only one component of the rest of the platform, the "*Sensing and Acting*" component, which belongs to the upper level: the "*Edge layer*". Once at this layer, the data coming from the system is analysed and computed by the components there, which is precisely why the platform uses edge computing technology. In fact, data acquired by the devices are analysed and processed in the layer as close as possible to the machines from which they originate. The processing can take place directly on the devices if they are equipped with sensors capable of processing data, or at the gateway to which the devices refer, i.e. a router, a switch, etc.

In support the edge layer, there is the "*Cloud*", where the heart of the platform resides. It contains the services, which provide assistance to the end user. They host the algorithms designed to perform the tasks for which the platform was created. The cloud also contains many of the components presented in section 2.1, which are essential for the smooth operation of the entire virtual system. Examples of these are the REST API interface, the Message broker and the Message bus & communication management, with the functions already described in the aforementioned paragraph.

Finally, on the left-hand side of the image we find the application layer, i.e. all those that are defined as the front face of the IoT solution. Examples of these are all the operator panels, desktop based or mobile based. Their main purpose is to present to the operator the final data coming from the platform, so that the latter can monitor the system and interact with it. Many applications also have the ability to receive specific input from the operator and allow the exchange of data at the interface level with other platforms and applications, or to forward the data received to the cloud for analysis, if the platform requires it. [5]

In order to describe its operation in more detail, it is necessary to proceed in order, following the flow of data and focusing on what and how each component present performs its function.

3.2.1 Data flow in the edge layer

As previously stated, the data comes from the plant and is acquired by the "Sensing and Acting" component. This component has the function of acquiring data from the plant and transforming the "physical" parameters, belonging to the analogue world, into electrical signals and digital data for future computation. If this conversion did not take place, the cloud would not be able to understand and analyse them. It also performs a further function, that of converting the digital data coming from the platform into physical signals that can be interpreted by the actuators and sensors installed on the devices that compose the physical system.

Since the data coming from the system originate from different sources, they have different type of formats. The task of the platform is to unify all these data into a common data model. In order to enable the collection of interoperable data, it is necessary to use data models specific for predictive maintenance applications. The digital models that enable different Machine-Learning algorithms to leverage the data into a uniform format, are essentially the following:

- Data Source Definition (DSD): this data model is useful to define data source’s properties (properties of a sensor is different respect to an autonomous device).
- Data Interface Specification (DI): Each sensor is associated with an DI that provides information on how to access its data and connect to the device, including details about port, address and network protocol.
- Data Kind (DK): Describes the data source’s semantics, in virtually way.
- Data Source Manifest (DSM): Associated with the previous digital models, contributes to specify a data source’s specific instance. Therefore, in order to

represent all data sources belonging to the physical system multiple manifests are needed.

- Observation: identifies a group of data generated from a data source's instance. It is a collection of information able to identify a dataset. Looking at figure 22 it is possible to understand this concept: associated with an observation, two timestamps store in which instant the data is acquired and when it is available to be captured by the platform. Moreover, it also contains information such as the DSM address of the data to which it refers and the DK entity, mandatory to describe the value type of the data to which the observation refer to. Lastly, a linked location attribute specifies the virtual or physical placement of the data source.

  So, the observation is a collection of information of a single data or a dataset, able to provide any detail about them. It refers to all the data into the platform, even including those resulting from analytical processing.

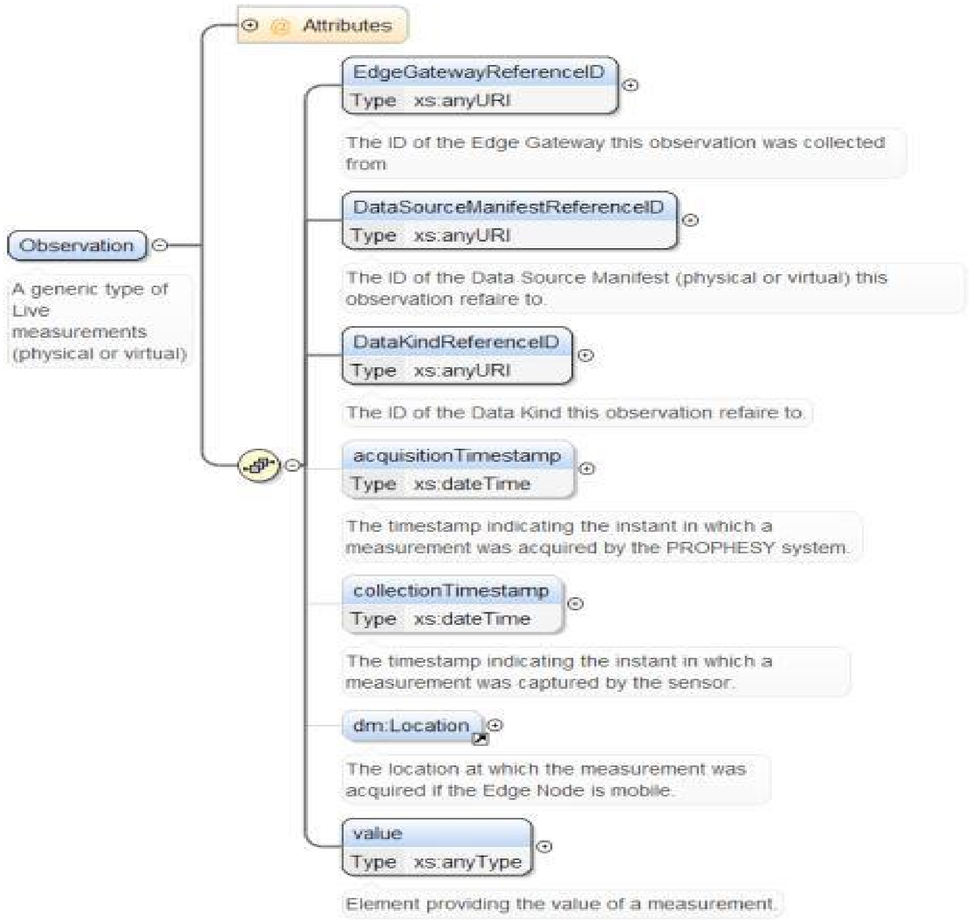

Attributes
Observation
A generic type of Live measurements (physical or virtual)
EdgeGatewayReferenceID
Type xs:anyURI
The ID of the Edge Gateway this observation was collected from
DataSourceManifestReferenceID
Type xs:anyURI
The ID of the Data Source Manifest (physical or virtual) this observation refaire to.
DataKindReferenceID
Type xs:anyURI
The ID of the Data Kind this observation refaire to.
acquisitionTimestamp
Type xs:dateTime
The timestamp indicating the instant in which a measurement was acquired by the PROPHESY system.
collectionTimestamp
Type xs:dateTime
The timestamp indicating the instant in which a measurement was captured by the sensor.
dm:Location
The location at which the measurement was acquired if the Edge Node is mobile.
value
Type xs:anyType
Element providing the value of a measurement.

Figure 22- Information contained into an observation: Digital Models Schemas

- Edge Gateway: refers to the modelling of an edge gateway belonging to a platform that exploits specific edge computing technology. From the deployment point of view, all data sources are associated with an edge gateways, not only logically but also physically, as each gateway deployed at a station has the task of managing and computing the data coming from the sources it is associated with.

The digital models are also useful to specify analytics entities related to the processing of data from each data sources. These entities refer to how data have to be computed, specifying the related processing functions to be applied, enabling the implementation of dedicated techniques with the ultimate aim of correctly managing and configuring the analytical functions to be used according to the data to be analysed. Analytical entities include:

- Analytics Processor Definition (APD): identifies which processing function to use for the referenced data source. There are three types of functions and they can be combined to create different workflows, depending on how the data has to be processed. The three types are functions that have the task of pre-processing the data, functions that have to perform analyses on the data and functions with the aim of storing the data coming from the computations.
- Analytics Processor Manifest (APM): the processor and the related programming function is specified through the manifest. It indicates the format of the implementation of the program that has to be applied for data computation (i.e. Java language, C language, and so on).
- Analytics Orchestrator Manifest (AM): A whole analytic workflow is described by this entity. It specifies the combination of APMs that define an entire analytics task, also underlining how the involved edge gateways have to operate during the computation of their data sources.

A similar argument can be made for the choice of configuration regarding analytics jobs, i.e. how the processor should operate for data processing. Again, the platform provides three entities: the Processor Definition (PD) file; the Processor Manifest (PM) file; the Processor Orchestrator (PO) file, with similar functionalities to those seen above.

The analytical entities described above work together to guide the configuration of the processor and define the workflow for the computation of incoming data. In particular, under the guidance of the processor orchestrator, the processor-engine is configured in run time in order to apply the right algorithms according to the type of data to be analysed at that precise moment. For this to happen, the format required by the tools to be used is defined in the analytics processor manifest, while the type of processor configuration is described by the processor manifest. Once the system is configured, computation is carried out in the terms described by the processor orchestrator, while the tools to be used are chosen according to the indications provided by the AM, depending on the workflow under examination. [19]

However, before the processor is called upon to compute the data, certain streamlining operations are carried out on them, so that the system is not overburdened. First, the data acquired at run time are temporarily stored in the "*Time-series Data Storage*". Then, the first operation that the platform performs on the stored data is pre-processing. This operation is still carried out in the edge gateway, by means of the "*feature engineering component*", following very precise rules, defined a priori by the "*pre-processing rules*". The task of feature engineering is therefore to streamline the amount of data present in the "*Time-series Data Storage*", replacing it with more significant data resulting from an initial computation of the first ones. This component is in fact able to extrapolate the average value from a set of data, the RMS value, the deviance and so on. [20]

Other operations carried out during the pre-processing phase are denoising, outlier removing and data-cleaning. In order to choose which data to remove, it is necessary to use the method based on the cycle length decile. Experts have noticed that some cycles, especially those belonging to the first or last decile, do not provide reliable measurements as they are far from real cases. Some of them derive for example from cycles used to test the correct functioning of the production process and therefore have to be removed in order not to alter the values resulting from the computation of the collected data. Once removed, these data are replaced so that the fixed-time structure is still respected. This alignment task can be performed in several ways. One example is to repeat the last value of the cycle until filling the cycle time slot, thus keeping the structure of the collected data compliant with the previous ones. The pre-processing task ends with two additional and sequential steps, called "*Statistics Computation*" and "*Smart Data computation*". These phases are essential, since the data, coming from sensor in raw time-series, need to be converted in a time-independent feature set, in order to be computed by the following components. The Statistics Computation takes place from the beginning, when contiguous portion of data are generated by splitting the acquired time series. The size of these portions is previously decided and can also be different each other, but the data computation is easier in the first case. Then, the extrapolation of statistical features previously described is performed: for each portion, statistics parameters (such as quartiles, min, max, standard deviation, mean and so on) are obtained, in order to summarize the original data. Not all these data are retained, but the less significant ones are discarded following two principles: the first is called "*multicollinearity-based*" and is based on removing those values that can be predicted according to multiple regression models on the basis of the other attributes. The second technique is called "*correlation-based*", which removes all the most correlated values. The parameter on which the correlation is based in this case is the average of all other attributes involved. Now, the second step is to aggregate the data from Statistics Computation into groups and compute them in order to extract new data again, to further reduce the number of data to be stored. Smart data computing is necessary because the data extrapolated from a single cycle cover a too narrow time window to express the degradation of the phenomena under investigation and make a reliable prediction possible. Therefore, by forming data sets belonging to different cycles it is possible to raise the prediction horizon of the degradation phenomenon and obtain statistical data related to a multi-cycle time window. The procedure that allows the degradation of statistical characteristics to be noticeably highlighted is based on the application of linear regression of the data belonging to the macro-groups formed previously, evaluating the slope of the resulting line and the coefficients that make it up. For each group, the mean, the minimum and maximum value and the standard deviation are then calculated and recorded. [18]

Two arrows leave the feature engineering in different directions, the first heading for the cloud, where the largest database of the entire platform is located. Here, all the data acquired during all the cycles of the machine are stored, already pre-processed. By analysing and reprocessing this data, the cloud is able to create new models that provide increasingly accurate estimates of the analysed process. The following section is dedicated to its discussion. The second arrow continues within the edge layer and expresses the relationship between the feature engineering and the "*CPU processor engine*" (CPU-PE) component. The latter is one of the two parts that make up the rule engine, the executive component of the whole platform. The other part, called "*PdM processor engine*" (PDM-PE), is located in the cloud and performs essentially the same function but at a different level. They constitute the "Machine-Learning (ML) Toolkit", a set of algorithms for the actual computation of the collected data. The purpose of them is different in that the CPU-PE is responsible for generating the final data that will be viewable by the user, using a model coming from the cloud, while the PDM-PE is responsible, among other tasks, for generating the model used by the CPU-PE. The data and the way they are processed are also different: the CPU-PE acquires the data from the feature engineering component, then analyses only the new data just acquired and allocated in the edge layer in run-time, while the PDM-PE computes all the data acquired up to that moment, stored in the cloud. Observing the arrows in the image it is in fact evident that the CPU-PE has as input the feature engineering, while the PDM-PE receives the information directly from the data storage. Since the PDM-PE has to process much more data and has to provide results from more complex and laborious analyses, it is therefore obvious that the way in which the data is processed must also be different: The CPU-PE uses HF-ML (High-Frequency Machine Learning) algorithms as it has to process smaller and less complex data streams, whereas the PDM-PE uses LF-ML (Low-Frequency Machine Learning) algorithms, for the reasons stated above. [19]

To better understand the potential of the PDM-PE and the substantial difference between the two, it is necessary to think that if the plant were to be expanded with the addition of other production lines and other machinery of a different nature, it would be necessary to add other edge gateway devices, each hosting a specific CPU-PE dedicated to it, but the PDM one, allocated in the single cloud would be the same and would have the task of managing all the data coming from all the edge gateways devices. The appearance of the platform would be as follows:

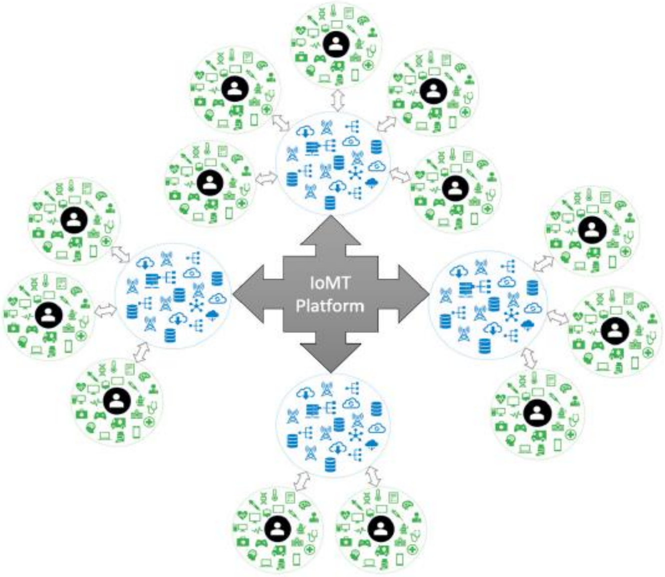


Figure 23- Edge computing IoT platform [21]

The resulting architecture therefore appears decentralised, as the first data computation takes place close to the place where the data are acquired. Each blue circle represents the edges, containing a database for storing data from the end devices (green circles) and a processor capable of analysing them. The different edge layers exchange data with the cloud, located in the centre of the image, so that the latter can carry out more complex processing and coordinate the entire system. [21]

The results obtained by the processor are stored in the last component in this layer. This is once again a data storage. However, the latter stores the final data on which no further processing will take place. Nevertheless, this data must be passed to the cloud, as it can only be viewed through a dedicated service. The detailed discussion of the service just mentioned will be one of the topics of the next paragraph.

The following image summarizes the data stream described into the edge layer.

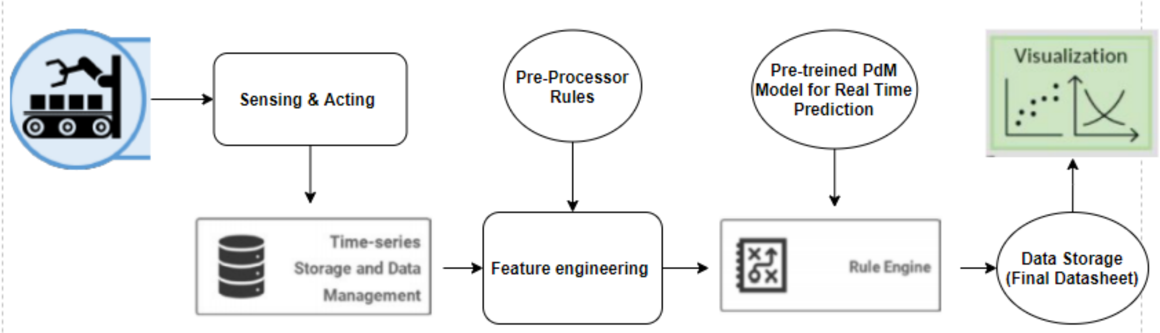

Sensing & Acting
Pre-Processor Rules
Pre-treined PdM Model for Real Time Prediction
Visualization
Time-series Storage and Data Management
Feature engineering
Rule Engine
Data Storage (Final Datasheet)

Figure 24 - Data flow in the edge layer

3.2.2 Data flow in the Cloud

As declared several times in the previous paragraphs, the cloud is the orchestrator of the platform. This is due to the fact that it hosts the virtual components and specific software that allow it to function properly. The presence of the REST API interface and the message broker, make possible the communication of the other components as they translate for each one the messages coming from the sender according to the formal messaging protocol of the receiver. Without this operation, the components could not exchange information accordingly. [5]

Given their particular function, they are tightly bi-univocally linked to all components in the platform, whether belonging to the cloud or to all other layers. Therefore, it has been chosen to represent them in the image 25 in a single block (the red one), omitting most of the links, in order to make the scheme clearer and easier to understand.

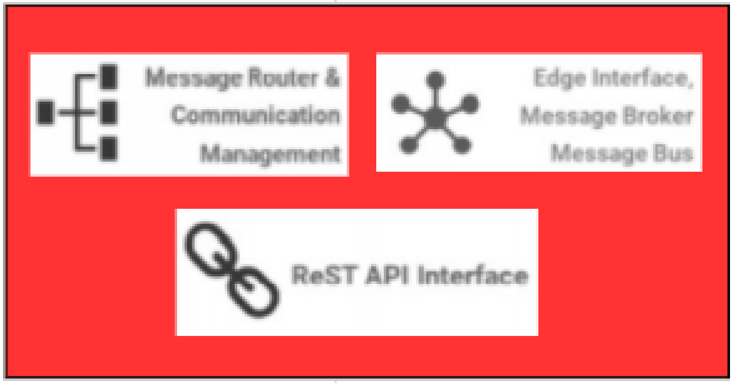

Message Router & Communication Management
Edge Interface, Message Broker Message Bus
ReST API Interface

Figure 25 - Main blocks of the cloud

It is also through the cloud that the platform is able to manage complex analysis and decide which decisions to take autonomously optimizing the entire process. To demonstrate the truthfulness of this statement, the flow followed by the data in the cloud and how the components present interact with it is described below, in a similar way to what has been done for the edge layer.

Information coming from the edge layer are storage in a huge database, hosting pre-processed data. Here, the data are acquired by the PDM – processor engine as seen before, and further computation are applied. In order for the platform to achieve the goals for which it was created, the processor works in collaboration with four applications integrated in the cloud, called services, each with a specific function. Also in this case, the services are collected in a single block, the yellow one, to make the design cleaner.

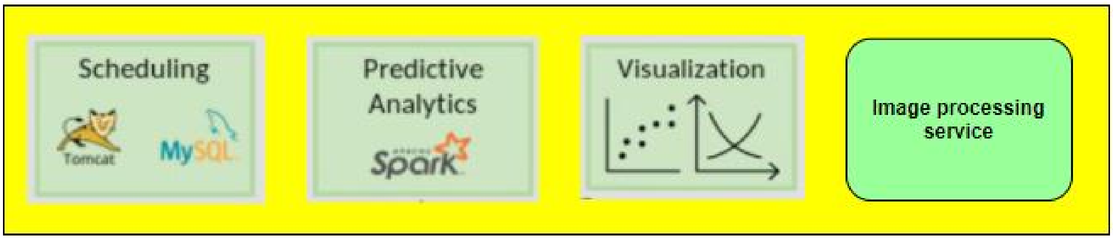


Figure 26 – Services of the platform

The first service analysed is the "*Predictive Analytic Service*", already briefly mentioned before. The purpose of this application is to generate a model, on the basis of the history of the data acquired up to that moment, able to predict the operation of the system and to estimate in the most accurate way the future course of the process in examination.

The following image clarifies its operation, analysing the data stream.

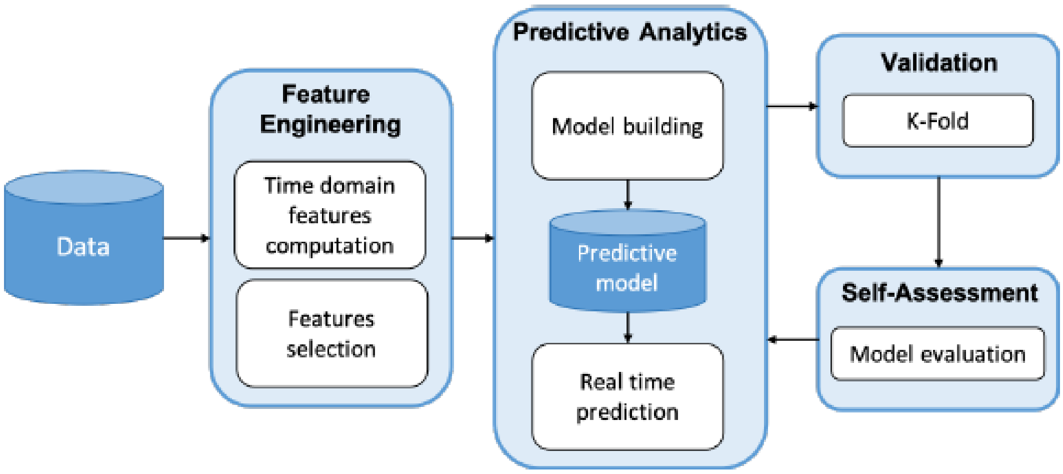


Figure 27 - Predictive Analytic Service architecture [20]

The reason why the pre-processing step is necessary is obvious in this context. Without the feature engineering, the service would receive a mass of attributes too large to process, containing data closely related to others, to be redundant and cause possible noise in the model building phase. Therefore, in addition to the phase of data cleaning and extrapolation of statistics features (identified in the image with the name of "*Time domain feature computation*"), the component also operates a feature selection phase, implementing the so-called "*Smart Computation*", seen in the previous paragraph. [20]

The first block representing the service is the "*Predictive Analytics*", in which a prediction model is built starting from the historical data collected in the data storage hosted in the cloud. Hence, the aim of the service is to build a predictive model, able to predict future events from the analysis of the new incoming data sensed by the plant in real time prediction phase. In the building model phase, algorithms are applied to obtain the better solution, able to relate with higher accuracy the events to predict with the input signals coming from the entire process involving the devices under examination. More than one model is created during this phase and each of them are validated through the "*Validation*" block. The scope is to understand which model best meets the requirements of the task to be performed by the machine at that precise moment. The validation phase can also be computed by different algorithms such as the "Stratified K-Fold Cross" validation technique. When the training phase is completed and the more suitable model is selected, the latter is sent to the edge layer, where it is executed by the CPU processor engine in order to fulfil the current tasks, obtaining results easily interpreted by the operator on the work station and in short time, since near to the edge of the system. Instead, the other models are not discarded but stored in a dedicated component called "*Model Repository*", since as they may be useful in the future. All the operation described before are executed thanks to the PDM processor engine, that it is involved every time an algorithm is applied, demonstrating that the processor and all services in the cloud are always working together. The last block of the service is called "*Self-assessment*". Here, a current predictive model is continuously evaluated through retraining, using the same dataset, but updated with new data coming from the physical system. This operation is important as it ensures that the model used is always suitable and returns outputs in line with the events that are happening at run time. It is often the case that with the addition of new machinery, changes to the production line or the degradation of the system or part of it due to time, the conditions under which the model is applied change drastically leading to a consequent degradation of the performance of the model itself. Then, the current model has to be discarded and a new model must be deployed in the edge layer, so that predictions of future events deviate as little as possible from reality. [20]

Therefore, the building model phase and the validation of the extrapolated models on the basis of the historical dataset is implemented in the cloud, but the application of those models takes place in the edge layer, where real time prediction is computed, using only the new sensed data acquired in real time. This feature once again underlines the nature of the platform based on edge computing technology. Moreover, it is now clear that the decision-making nature of the platform consists in the choice of the best possible model, capable of obtaining the prediction of the most suitable results on the basis of the current circumstances in which the system is.

The ability to make decisions autonomously in a very short time is also found in the "*Scheduling Service*". Thanks to the service mentioned before, the platform is able to change the current production plan, reorganising the numerous tasks of the production system, including maintenance operations by operators, in order to reduce unplanned downtime and optimise production execution times. The service receives as input the results obtained by the Predictive Analytics service and works as a client-server model. The client side includes as input, in addition to the results achieved by the predictive model, all devices belonging to the monitored equipment, the planned tasks assigned to them, the time-slot dedicated for plant maintenance and the experience level of the operators. The server hosts a framework of decision making, capable to generate several scheduling configurations different with respect to the current one, to assign to each of them a score indicating the level of suitability, valid for that precise moment under those circumstances, and to select the configuration with the higher rank. Therefore, the use of the scheduling service helps to prevent a sudden failure of a machine or device in the system, which can lead to a decrease in productivity, by reconfiguring the existing tasks and adding those aimed at system maintenance. [20]

Once the data is processed by the platform, it is necessary for the cloud to transmit the results of its analysis in a way that is easily understood by users. In order to perform this functionality, the "*Visualization Service*" is hosted into the cloud and communicates with the upper layer of the platform, where a panel operator is included. Thanks to the service, engineers are allowed to evaluate the status of the system and the devices and perform considerations about the efficiency of the productivity, observing the models generated and their outputs. The maintenance operators are instead supported in their maintenance activities. The presentation of the system's status can be in different way such as presenting the most significant outputs coming from the platform's computation or for example showing 3D images of the most damaged components of the system that need to be replaced. Hence, it is easy to deduce that services also communicate with each other as well as with other devices belonging to the platform, according to the REST APIs, in order to properly perform their operations. [20]

In addition to displaying the results and any alarms, the operator panel can be used by the operator himself to load significant images from outside of the platform. Loaded, the images are sent into the cloud, where algorithms are recalled in order to extrapolate information on them. The algorithms are belonging to the "*Image Processing Service*" and can be of different format such as Matlab/Phyton and so on. Hence, this service communicates with different direction with the platform respect to the other services: it receives inputs from the application layer and sent his output to components into the cloud. In fact, the data obtained from the processing of the images are stored in the database hosting the pre-processed data, and will be used for the building of new models by the predictive analytics service. If there is a need to keep track of the images acquired, a “*Image Repository*” component could easily be integrated into the cloud to keep track of the visual representation of the data obtained from them.

Within the platform there is in fact a network of repositories aimed at storing useful data for the services present and for the numerous operations that can be carried out by the platform. An example of this is the “*Model Repository*”, already seen above, where all the models created by the predictive analytics service and used in the edge layer are stored, as they may be useful again to the system. Another example is the “*Device Registry*”, a database where all the devices belonging to the physical system are registered with information on their identification and operation attached. The addition of a possible “*Metadata Repository*” is also envisaged, aimed at storing metadata containing additional information on the collected and stored data. [19]

There is still one component within the cloud that can provide very powerful functionality to the platform: it is the “*Virtual folder synchronization”* component. It is capable of connect the platform with other platforms and collect data and measurement coming from their systems. Obviously, these data are in different format and have to be pre-processed in order to be translated in the uniform format of the others data in the platform. The Virtual folder synchronization component is therefore connected to the time-series data storage. [19]

Chapter 4: Case study: resistance spot welding

In this chapter, the resistance spot welding process will be presented, with particular emphasis on the parameters that most influence the quality of the welded spot. It will describe in detail how the recrystallisation of the melt occurs and what hazards can be encountered during the solidification of the joint. In the last section the phenomenon of pitting will be presented and how it affects the wear of the electrodes in resistance spot welding.

4.1 Introduction to the process

Resistance spot welding (RSW) is an electrical resistance welding (ERW) useful to weld metal sheets of various type and composition. As a typical electrical resistance welding, the welding process consists of putting in contact metal parts and melting the metal in the area of the contact, creating a permanent joint. The joint is created as a consequence of joule effect, resulting from the heat obtained from resistance to electrical current. Indeed, the entire process is allowed by using two shaped electrodes, typically in copper alloy, that performing a pression in the limited area, concentrate welding current into a small spot. This process simultaneously clamps the sheets together due to the heat generated by the flow of the current through electrodes and metal sheets. Due to Joule effect, the generated heat is proportional to the resistance of the spot, following the basic formula:

$$E = Ri^2t \qquad (1)$$

where *E* is the transferred heat, in *Juole*; *R* is the local resistance, in *ohm*; *i* the current's intensity, in *Ampere*; *t* the time interval in which the current flows through the spot, in *second*. [22]

The factor *R* is composed by three terms: the first of them, takes into account the contact resistance between the electrode (of copper alloy) and the two metal sheets (typically in iron alloy); the second term, the resistance of the iron parts to weld and the third the contact resistance between the two iron sheets. The biggest one, is the third one, because the elasticity module of the copper is smaller than the iron, so the contact areas between electrode-sheet is bigger than the contact area sheet-sheet at microscopic level. A smaller contact area means greater resistance to the passage of current. As a consequence, the maximum heat is generated between the two iron areas in contact, creating a limited spot during the welding. Now, it is clear that the amount of the delivered heat is influenced by the intensity, the duration of the current's flow and the resistance of the parts involved and playing with these factors it is possible produce the right quantitative of the energy to match the sheet's material properties, thickness, electrode's type and size, in order to obtain reliable spot welds. A more in-depth analysis of these topics will be discussed in the following sections. [23]

4.2 Influence of welding parameters on welded joint

The welding parameters play an essential role during resistance spot welding. The quality of the spot weld must satisfy specific requirements demanded by costumers for which the final products are intended. Examples of these are high strength of sheets, stiffness of weld joint subjected by external load, ability to absorb impacts, corrosion resistance and so on. In order to maintain the specifications above, some investigations were done by the manufacturers. The following discussion is intended to show how a welding parameter's variation can affect the spot weld properties. Therefore, the influence of the three main factors responsible from heat generation during RSW will be investigated: the resistance of the conductor, the intensity and duration of the current. The treatment will be accompanied by studies and experimental tests to proof the assumptions done.

4.2.1 Electrical resistance

As seen in the previous section, the electrical resistance can be divided in three factors. However, a deep analysis can be done considering the following picture (Fig. 28), where a more precise division of the resistances in play is shown.

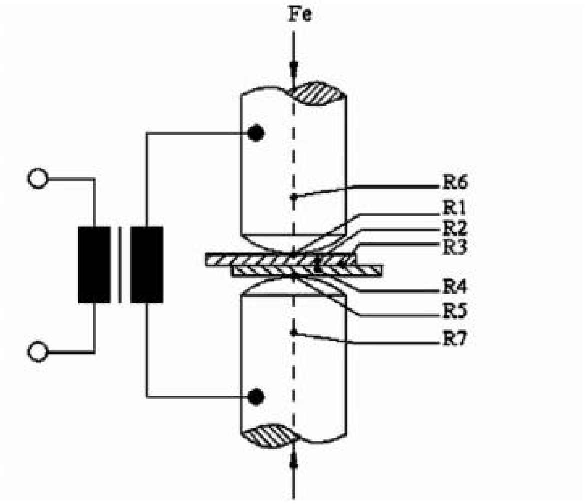


Figure 28- Electrical resistance in RSW [24]

In a weld, there are actually seven resistances connected in series and all of them are component of the *R,* variable of the equation (1). Depending on their peculiarity, they can be grouped in three sets: *R2*, *R4*, *R6* and *R7* are the so called "*bulk resistances*" and are the material resistances; *R1* and *R5* are the "*contact resistance*" between electrodes and workpiece; R3, the most important one, is the resistance at the sheet faying interface. The value of this quantity is not constant but depends on the surface condition of the metal sheets and the electrode involved during the welding. An easy weldability will be possible as much as the value of *R3* will be high. The other resistance values are also not constant: contact resistances are function of temperature and pressure, bulk resistances are sensitive to temperature only. [24]

In the following scheme, the dependence of bulk resistivity on the temperature is graphically represented.

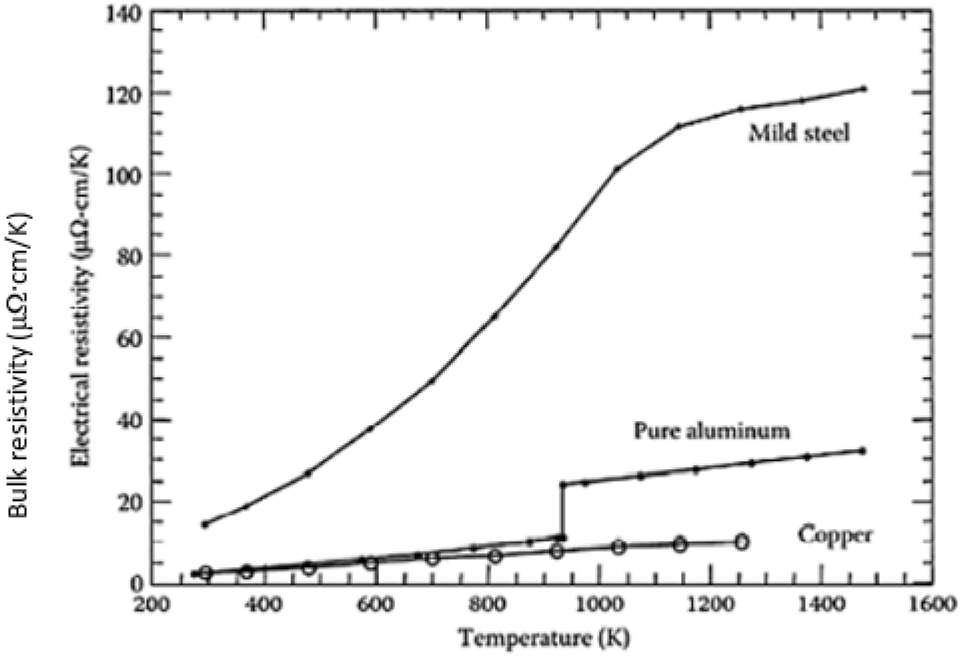


Figure 29- Bulk resistances depending on Temperature [25]

The study was conducted on three different materials, the most used in the industry. All of them have an uphill trend, but with different rates. Increasing the temperature, the bulk resistivity of the mild steel is more sensitive than pure aluminum and copper, for each range of temperature. For this reason, in a welding process, copper, or an alloy of it, is used as electrodes material, instead meld steel as workpiece. Indeed, the comparison between the resistivity values of the materials above, suggests that the most quantity of heat is exhaled in the workpiece than in the electrode when an electric current is applied. Moreover, the electrodes are usually water cooled, in order to further reduce the effects of heat on them.

However, in motor-car manufactory, the use of aluminum is more required with respect to meld steel for its lightness, high strength to weight ratio and other performances. Although as highlighted by the graph above, the welding pure aluminum using copper as electrodes is difficult, due to the trend of the pure aluminum resistivity too close to that of copper. To solve this problem, various alloys of aluminum are more used as workpiece, being the latter as sensitive to temperature as the meld steel. [25]

About the contact resistance, a more detailed analysis will be presented, trying to explain why it plays such an importance role in resistance welding. It provides the greatest contribution to the total resistance's determination and is made up of two components:

constriction resistance and film resistance. The first one comes from the actual material volume locally used at the interface, where the current flow lines are forced to cross the interface only through the separated conducting spot in real contact. Indeed, due to the pressure of the electrode, the asperities of the two sheets resulting thanks to the roughness of the material, are squeezed and smashed, resulting in a reduction of the real contact area, less than expected. Instead, the second one, is due to the surface condition of the interface. The presence of oxides, oil, dirt, paints, water vapor, etc. affects the conductivity of the surface, causing a change in the resistance.

A study was conducted on a Gleeble machine, in order to analyze the influence of electrodes' pressure and temperature on contact resistance in three different welding materials: carbon steel, stainless steel and aluminum. The Gleeble machine is a dynamic testing system that simulate a various thermal and mechanical process thanks to a prescribed program able to measure and store a huge number of data, coming from parameters of interest. The test specimens, circular cylinders with size of the order of millimeters, are heated and loaded by the machine in order to investigate their contact resistance in a wide range of temperature (from room temperature to melting point) and a typical range of pressure used during a real resistance welding process. The Gleeble system is setup as in Fig. 30.

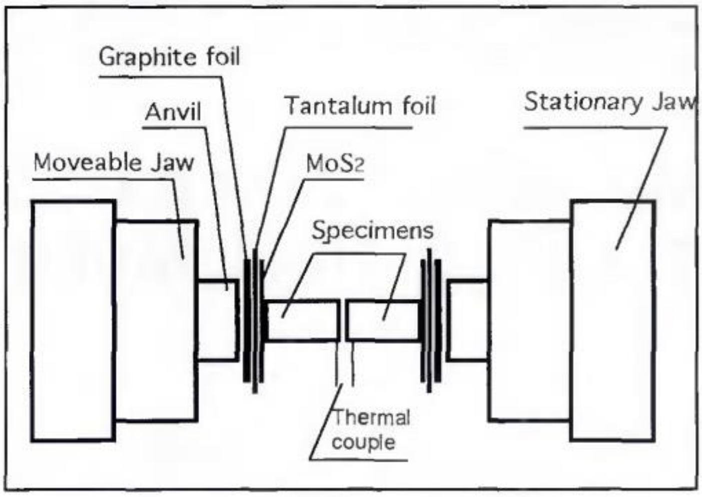


Figure 30- Gleeble system [26]

The system is formed by two anvils, one fixed to the movable jaw and the other fixed to the stationary one. Between the specimens and anvils, the presence of different plates ensures accuracy in temperature measurements: tantalum foils are placed for thermal insulation, graphite foils and molybdenum disulphide powder for friction reduction.

Moreover, the material of anvils and jaws is specially chosen to minimize the temperature gradient through the samples. The procedure of heating and loading the specimens during the test with desired values and related measures are made by a GPL (Gleeble Programming Language) program. Measurements of voltage drop on the faying surface and of current's flow allows the calculation of the contact resistance, thanks to the Ohm's law. Figure 31-33 shows the obtained results.

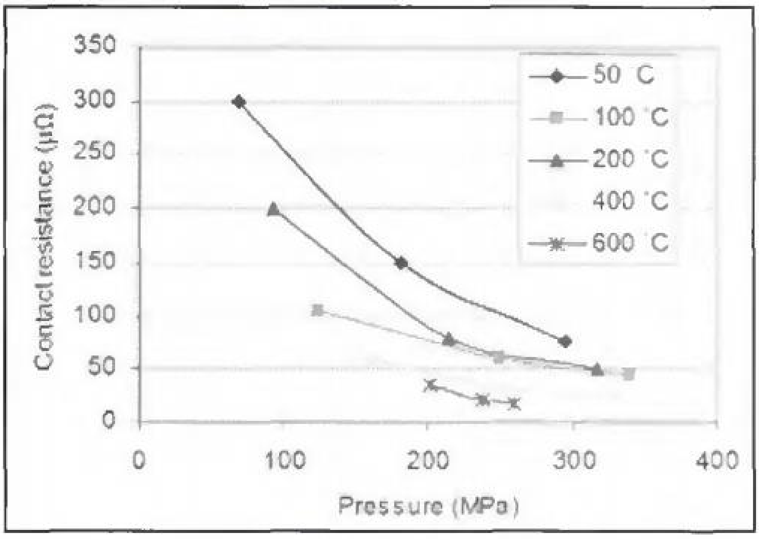


Figure 31-Contact resistance in carbon steel [26]

The graph above, illustrates the dependence of the carbon steel's contact resistance on pressure, for five different values of temperature. The influence of the pressure is higher for curves with low temperature and the slope of them is deeper at low pressures than at high ones. This latter behavior can be caused by an increasing of contact area and a rupture of the surface film, due to a considerable deformation, leaded by the increase of the pressure. However, in general, the increase of normal pressure involves a contact resistance's reduction. An interesting behavior is highlighted with the comparison of the curves: at 50°C the resistance is highest, decreases at 100°C but increases at 200°C and finally drops for highest values of temperature. This can be explained by the fact that the higher the temperature, the easier the film breaks.

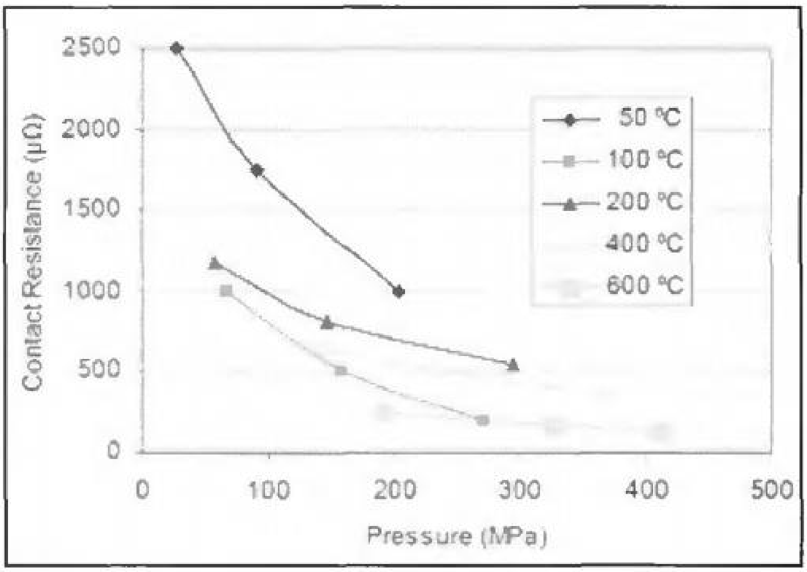


Figure 32-Contact resistance in stainless steel [26]

The influence of pressure and temperature in stainless steel is similar to the one shown before. However, due to the higher electrical resistivity, this material has in general value of the contact resistance much higher than that of the carbon steel. The figure 32 demonstrates these statements.

Instead, in the following graph it is possible to see the aluminum contact resistance behavior. The influence of temperature and pressure is as before, but now the smaller electrical resistivity leads to a lower value of contact resistance values with respect to carbon and stainless steel.

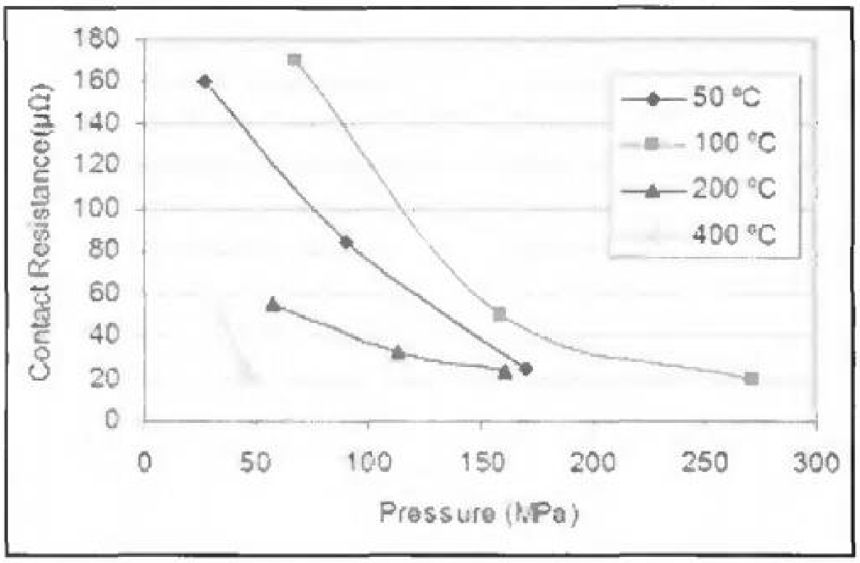


Figure 33-Contact resistance in aluminum [26]

Summarizing the results of the study, pressure acts in two ways, increasing the contact area and speeding up the break of the surface film, causing as result firstly the decrease of constriction resistance and then the decrease of the film resistance. When a more pressure is applied, the influence of these effects becomes negligible, since the contact area reaches the theoretical one and the film has been completely broken.

About the effect of the temperature, it is more complex, due to the influence of the electrical and mechanical properties of both films and materials involved: at higher temperature, an increase of the contact area, due to the metal's softening, reduces the constriction resistance, but the increase of the resistivity rises it. At the same time, some contaminants like water vapor can be burned and the breakage of the film layer can occur with greater ease, reducing the contact resistance; on the other hand growth of some layers can be caused by rising temperature (such as oxidation layer), increasing contact resistance. So final result depends on which effect prevails over the other. [26]

An example of how the total resistance's value (sum of bulk and contact resistances) changes due to joint effect of mechanical and electrical properties of both welding metal and film, is illustrated in the following figure.

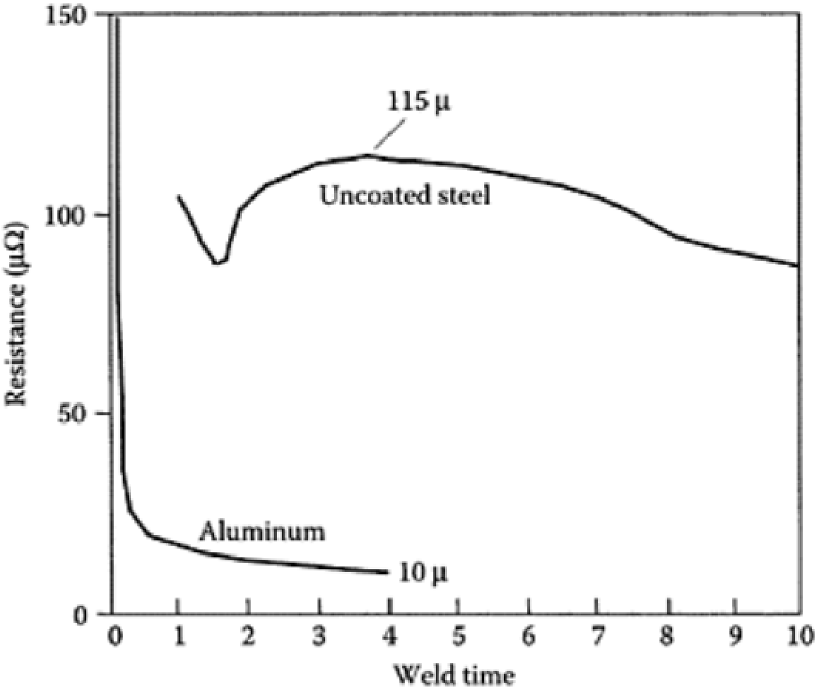


Figure 34-Comparison of steel and aluminum alloys' total resistance over the time [25]

The behavior of total resistances for steel and aluminum alloys are compared. The trend of aluminum alloy can be divided in two parts: the first one, with a steep drop in the first small range of time (from 0 to 0,5 cycles) and the second one, also with decreasing behavior but with a much smaller slope. The steep drop is due to the presence of Al_2O_3 layer on the surface, responsible of the great initial value of the total resistance. When the welding process starts, the current flow breaks the film and the resistance drastically drops. The second part of the behavior is due the joint effect described above: as the temperature rises, the metal becomes softer and leads to at larger contact area, resulting in a reduction resistivity. As a final result, the total resistance has a negative trend with small slope, despite the effect of the bulk resistivity, constantly increasing throughout all the process.

Looking at steel behavior, it is clear that its total resistance is significantly higher than the previous case, due to different thermal and electrical conductivities of the two metals, lower in steel. This trend has also a negative slope, but with the absence of a steep drop, due to a lower resistance of the surface layer in welding steel. When the layer is broken, the total resistance rises around 2 seconds. This peculiarity happens thanks to an increase of resistivity when metal is heated, this time no longer contrasted by the increase in the area of contact. Indeed, as showed in figure 29, the increasing of resistivity with temperature is more significant in steel with respect to the one in aluminum. Moreover, also in this case the metal is softer, but it still has the strength to oppose the pressure exerted by the electrodes. With more heat, the metal leaves this capability and the total resistance starts to decrease again. [25]

4.2.2 Electrical current

The welding parameter that most influences the quality of the point is the amount of current used during the process. It will see how the welding current affects the quality of weld joint at macro and micro structural level and the obtained spots will be tested to investigate how they react to shear and traction stress.

The intensity of the current is obtained from the circuit structure of electrical resistance spot welding machine. Being an alternative current, it results as:

$$I = \frac{U}{\left(\sqrt[2]{(R^2 + \omega^2 L^2)}\right)}$$

where R is the total resistance, L the total induction and ω the frequency of the circuit. The voltage of the system U is constant, so the only way to maintain constant the current, avoiding an excess of it that would cause cracks and voids during welding, is control the variation of resistance and induction of the circuit. The description of the following experiment has the aim to assess the influence of welding current on tensile-shear and tensile-peel strengths of the joint. Figure 35 shows size of specimens, couples of galvanized chromate steel sheets with 1.2 mm thickness (s).

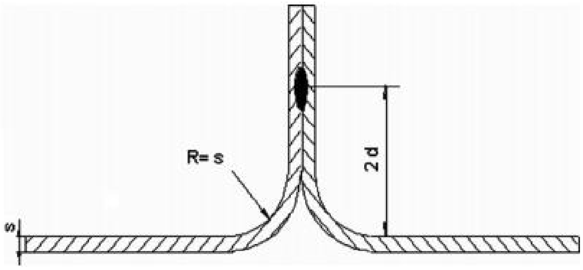


Figure 35 – Specimens sizes welded together [24]

All the other parameters are maintained constant (e.g. 6kN for electrode force), only current is increased in range from 4kA to 12 kA, with steps of 1kA at a time, and different value of weld period are applied and plotted: 5, 10, 12 and 15 cycles. The resulting weld joint are tested and three types of fracture are observed: separation, knotting and tearing. The first one appears for low value of welding current; instead, for high value of this parameter the breaking failures turn into knotting and tearing. The increase of welding current also causes an increase in the diameter of the welded point, badly affecting the welding quality: looking at the specimens, electrode marks are seen, due to an increase of melting material at the interface.

Figure 36 shows the dependence of tensile-shear force on welding current. The trend is positive until a maximum value (around 11kA for 5 cycle and 10 kA for the others), so the increase of current leads an increase of tensile-shear force. Reached the maximum point, the presence of electrode marks suggests an increase of nugget diameter, due to a melting of material interface between electrodes. The resulting effect badly afflicts the quality of the joint. Moreover, a reduction of cross section is recorded and tensile shear strength consequentially decreases too.

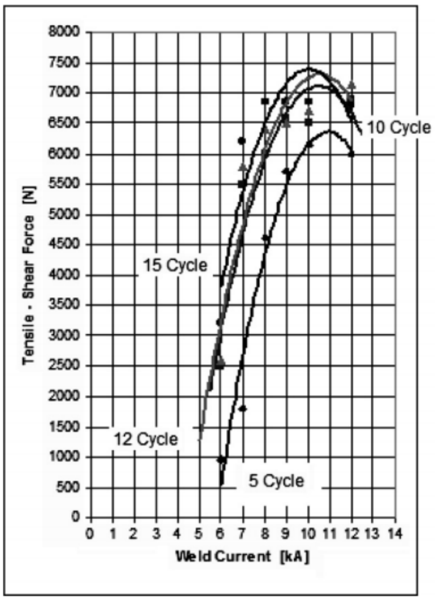


Figure 36-- Influence of welding current on tensile-shear force [24]

Similar behavior is shown in figure 37, where the dependence of tensile-peel strength is plotted. In this case the trend is also positive until a maximum value, then the strength starts to decrease. The best performance is reached in 10 cycles at 11 kA. However, the maximum value of tensile-peel force is lower than 2750 N, a value much smaller than the one of tensile-shear.

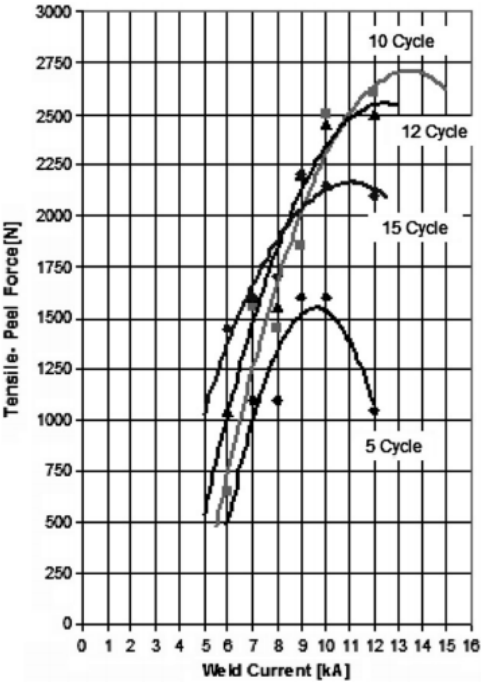


Figure 37- Influence of welding current on tensile-peel force [24]

In conclusion, the operating point of welding process has to be around the maximum value of the curve, where the weld joint has maximum strength to oppose to shear and tensile stress. Under the condition of the experiments, this means in a range of welding current from 10 kA to 12kA. An excess can affect badly the quality of welding joint due to an exaggerated depth of electrode indention into workpiece. [24]
In order to investigate how the welding current affects macrostructure and microstructure of a welding joint (and consequentially its mechanical properties such as hardness, etc.), a study is conducted on 16 mm thick cold-rolled of Dual phase steel, one of advanced new materials used in automotive industries belonging to the category of ADHSS (advance high strength steel). As in the previous experiment, all the parameters are kept constant, but the welding current moves in a range from 40 kA to 60 kA, in steps of 5 kA. For each step, a picture of cross-section is taken, in order to understand when defects occur. Figure 38 shows the two sheets not welded yet. So, intensity current of 40 kA is not sufficient to create a joint. Reached 45 kA, the welding process happens successfully, but defects are present (Fig. 39). The best solution is verified at 50 kA when proper penetration of sheets takes place, without defects (Fig.40). Increasing the current up to 50 kA, the joint integrity is compromised: undercut and metal deformations take place (respectively at 55 kA and 60 kA) as in figures 41-42.

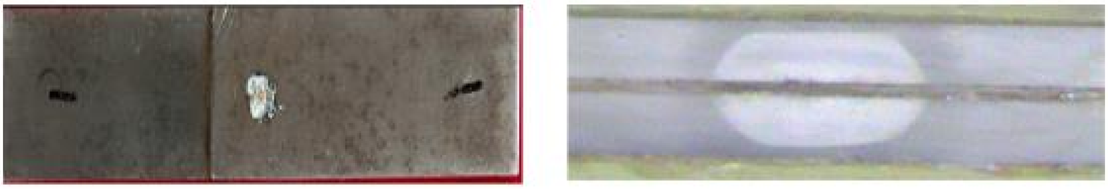


Figure 38- current at 40 kA, no welding [27]

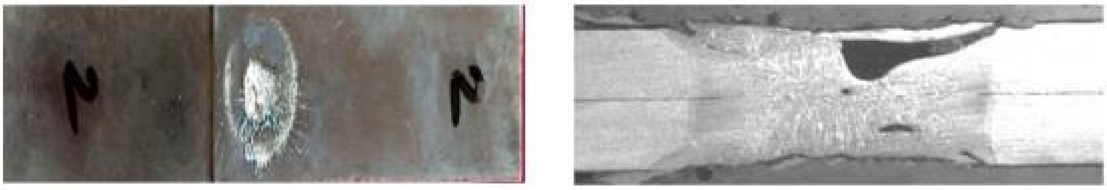


Figure 39- current 45 kA, welding with defects [27]

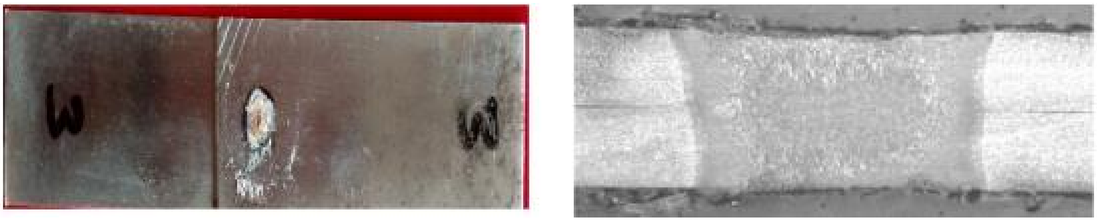


Figure 40- current at 50 kA, welding without defects [27]

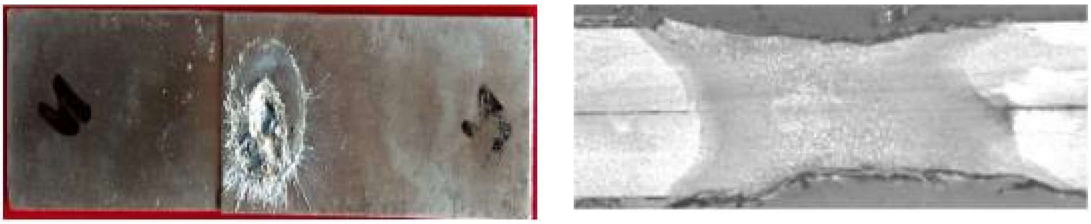


Figure 41- current at 55 kA, undercut phenomena [27]

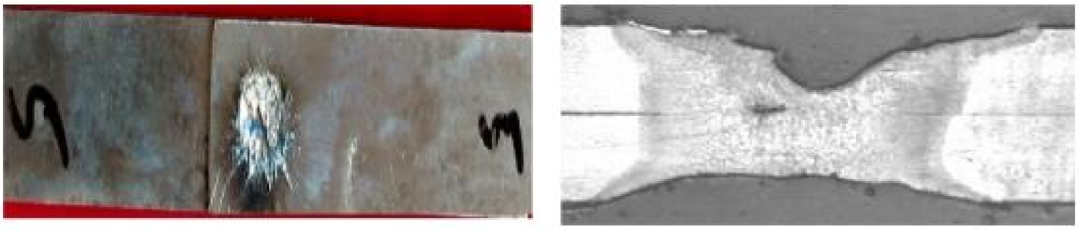


Figure 42- current at 60 kA, metal deformation [27]

At microstructural level, the heat affected zone (HAZ) and interfacial zone (IF) are analyzed to understand where cracks and voids take place and propagate in the nugget. Figure 43 illustrates how HAZ and IF crystallize in the best situation, when intensity current is about 50 kA. Solidification starts in the interfacial zone and end into the HAZ, with formation of grains oriented in dependence on the cooling rate. The weld failures are mostly located in IF. A deep analysis of this topic is treated in the next chapter.

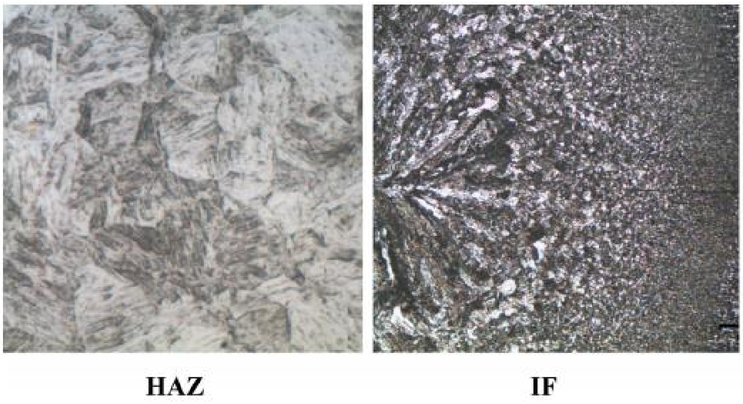


Figure 43-HAZ and IF in a weld spot joint using current at 50 kA [27]

After solidification, the specimens are tested in tensile-shear and cross-tensile and the crystal structure of nuggets is observed again. The nugget zone appears as in figure 44. The experiment confirm that tensile fractures are mostly located into the interfacial zone, starts from a point outside the weld nugget, at the edge of it and propagates along the workpiece until it ruptures.

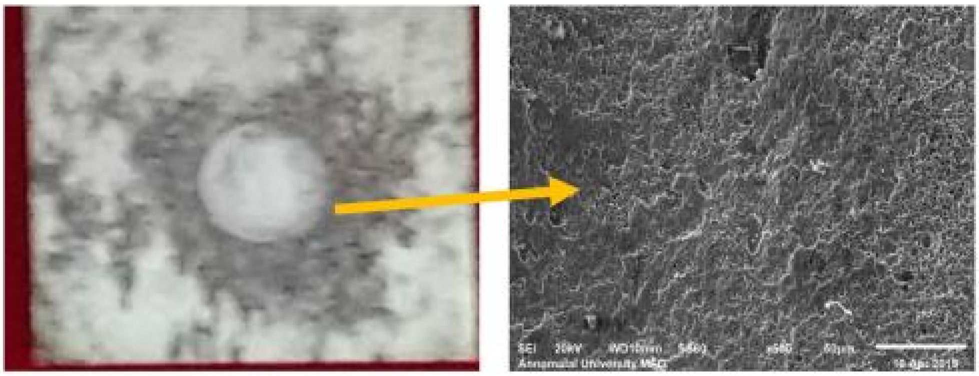


Figure 44-Fracture surface of nugget after tensile-shear test [27]

The last analyzed property is microhardness of specimens and its correlation with tensile-shear. The hardness is mapped in the figure below, dependently on distance from nugget center. The highest value (480 Hv) is located at the center and decreases as we get closer to the edge.

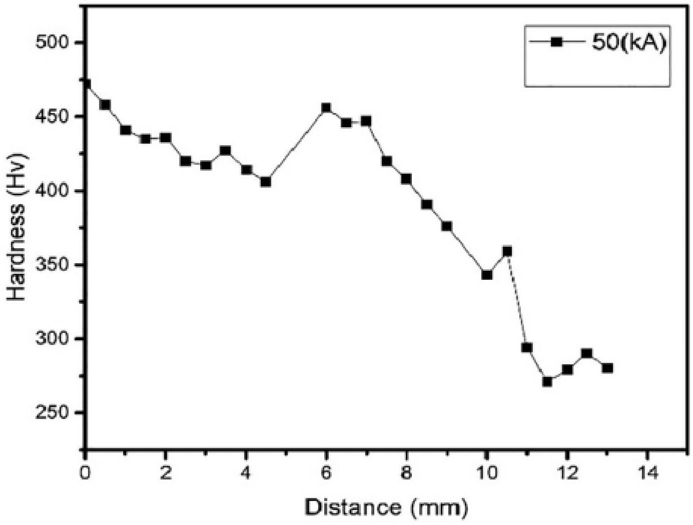


Figure 45- Hardness from nugget center, created using current at 50 kA [27]

4.2.3 Shunting

Shunting is a phenomenon that occurs when the welding points are too close to each other, due to high conductivity of some metals, such as aluminum. So, this effect is strictly dependent from bulk resistivity of sheets metal and can affect the subsequential welding, producing weld with no intended strength. If it occurs, the current destined to a specific welding spot divides, also flowing into previous spots close to the latter. As final result, the current flow in the intended spot may not be enough to produce the desired weld.
The overall effect can be shown in the following figure, where current flow lines are drawn. [25]

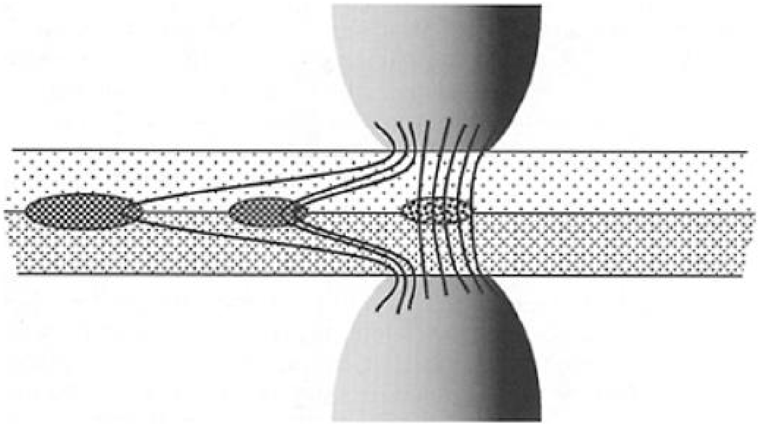


Figure 46-Electric current shunting in RSW [25]

4.2.4 Time effect

The last important welding parameter to take under control is the duration of the process. To realize a properly welding, electrodes have to be in contact with workpiece for a specific amount of time. If this time is too short, proper penetration between sheets does not occur; if it is too long, the heat generated could also spread to electrodes, causing overheating and consequent wear. The amount of time depends on various parameters of the workpiece: material composition, size, thickness, surface condition, etc.
A study to analyze the effect of time on welded joint is carried out. Welding parameters such as electrode type, form and force (6 kN), are kept constant, welding current and time are changed and the welded joint are tested in order to investigate tensile-shear and tensile-peel strength. The tested specimens are the same of the previous study (about

electrical effect): a chromate micro-alloyed steel and galvanized layer with thickness of 1.2 mm and 23 µm respectively. Welding current and time are set as before (time applied as 5,10,12 and 15 cycles and range of current from 5 kA to 12 kA with steps of 1 kA). Under these conditions it is possible to plot tensile-shear and tensile-peel strength in function of welding time, at different values of welding current. Results are illustrated in figures 47-48.

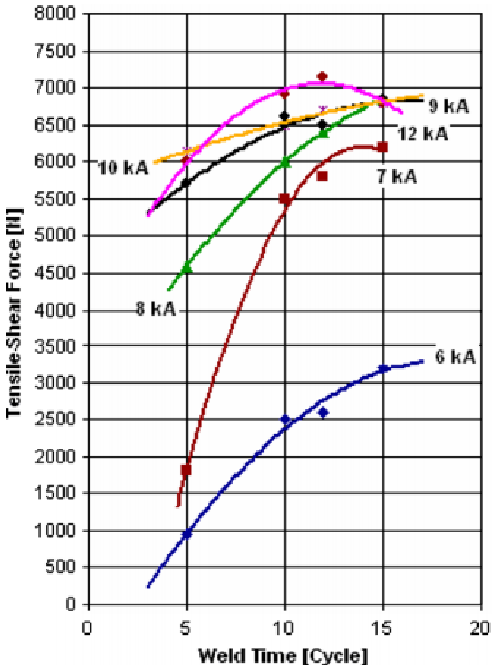


Figure 47 – Welding time effect on tensile-shear strength [28]

In general, the curves in figure above, show that increasing the welding time, tensile-shear strength is increased, caused by the expansion of the nugget, due to the increased amount of heat brought to the weld joint. Looking at the graph fixing the period of time, it possible to observe that tensile-shear strength also increases with bigger values of welding current, as we can expected by theoretical reasoning. This happens for each period, but in different ways: for period of 5 cycles, tensile-shear strength increases rapidly from 5 kA to 10 kA. After this value, the growth is negligible; for 10 and 12 periods, the same quickly growth is registered up to 8 kA, then the dependence variable's increase proceeds lower. Instead, different behavior takes place for 15 period, where the increase of tensile-shear strength of specimens is sharply up to 9 kA, reaching its maximum value, and then starts to decrease. This reduction is due to generation of excess heat, responsible of crack and void formations into the weld joint. Also in this case, three typey of failure can occur: separation, tearing and knotting.

The time effect on tensile-peel strength is similar to the one described above. The only difference is that the dependence variable starts to decrease also for curves of 7kA, 10kA and for 10,12 and 15 periods, increasing welding current, it does not always induce an increase in tensile-peel strength of specimens. The maximum value is reached at 12kA in 12 periods. The following figure summarizes the described results. [28]

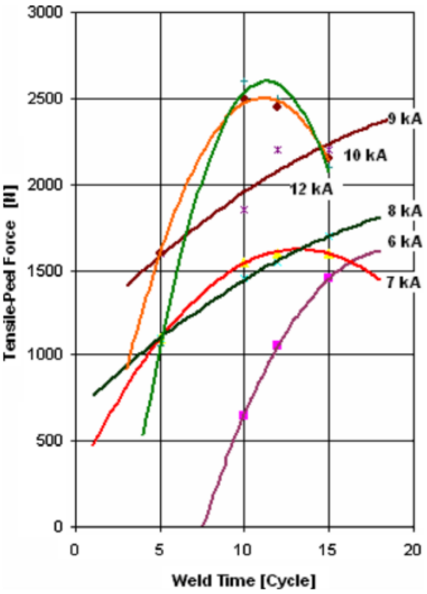


Figure 48– Welding time effect on tensile-peel strength [28]

4.3 Quality of welded joint

The welding process permits the transfer of much energy in a very short time (about 10-100 milliseconds), avoiding excessive heating of the remainder of the sheets. In order to clearly understand in which way the heat is transferred, it is necessary to study what happens in the crystal structure of the alloys involved after the melting metal, how the grains recrystallizes and how their influence the weld joint's strength.

4.3.1 Metallurgical principles

The aim of this chapter is to analyze the metallurgical principles governing the aspects of RSW and how they are critical in understanding the formation of the structure of welded joints. During welding, the solidification of the liquid part begins as usual with the nucleation and proceeds with the growth of the crystals. This last process is directly influenced by the heat dissipation into electrodes and metal sheets: the size, type and orientation of the generated crystals depends on the direction and the rate of cooling. If the rate is too quicky, the micro-segregation phenomena is arisen. In the spot weld, a stratify layers with different crystalline structure and chemical compositions take place, because an equilibrium composition distribution is not achieved due to an insufficient diffusion rate. So, at the end of this process, a different composition between the core and the outer layer are formed and the differences between them increases as much as increase, in the phase diagrams, the distance between the liquidus and solidus lines. The only ways to decrease these differences, is increasing the diffusion rate or increasing the time span useful to the solidification. In the alloys subjected to the welding process, the elements are involved in another phenomena, called segregation. It is the chemical-physical phenomenon where in a solid solution formed during solidification of a liquid one, the liquid component that solidifies with an upper melting point, solidifies with the native structure without the interference of the other component. So, due to the different melting point of the elements that formed the alloy, this process takes place as the solid-liquid interface advances into the liquid, and as result, the concentration of the element with lower melting point is increased in the remaining melt of alloying elements. [29]

A huge problem can occur in aluminum alloys, when at grain boundaries certain compounds, rejected from solid solution, are the last to solidify due to their lower melting temperatures. If the metal is stretched by thermal stresses during the process or by an external load, a crack can be generated at the grains surrounded by such liquid as the liquid has no stretch under these conditions. This problem can be solved in the RSW thanks to the right pression performed by electrodes.

Solidification starts at the boarders of the heat-affect zone (HAZ), where partially melted grains become nuclei of the new solid grains' growth, with columnar orientation. The central zone of the HAZ, solidifies last, on condition that the melt liquid volume is smaller than the solid volume surrounding it. This becomes the new site for the equiaxed grains' growth, oriented depending on the versus of the cooling rate. The energy dissipation, enabled in the liquid volume if the quantity of heat that goes out is bigger than the quantity that flows into, depends on the source of heating around it: the water-cooled electrodes and the cool metal sheets, that transfer the energy from the sides. Keeping in mind this influence, three different conditions can appear. The first one (Fig. 49) is the ideal one, in which the solidification occurs in uniform way from the electrode and sheets' sides. In this configuration, if cracks and voids are generated in the central portion of the nugget, they don't affect the performance of the welding.

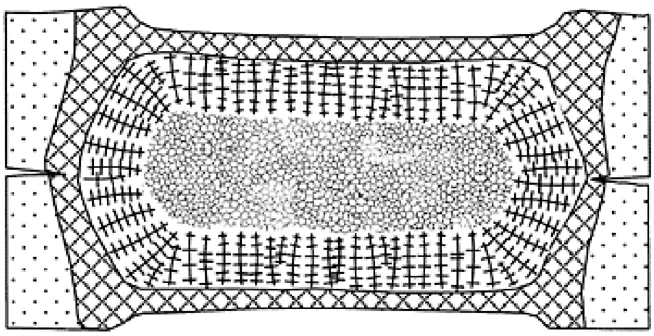


Figure 49-Grains in a uniform solidified joint [25]

A different structure is obtained when the electrodes are overcooled. Looking at figure 50, it can be seen long columnar grains in electrodes direction and smaller grains from the sides. This configuration takes places when the cooling rate is bigger in vertical direction, so due to faster solidification from top to the bottom, the last small liquid volume solidifies near to the original interface of the metal sheets. The situation is significantly dangerous because a deficit of volume is source of cracks between the sheets and due to the lower solidification rate at the lateral directions, they can be very close to the HAZ, since the reduced volume of liquid is located at the periphery of the spot weld.

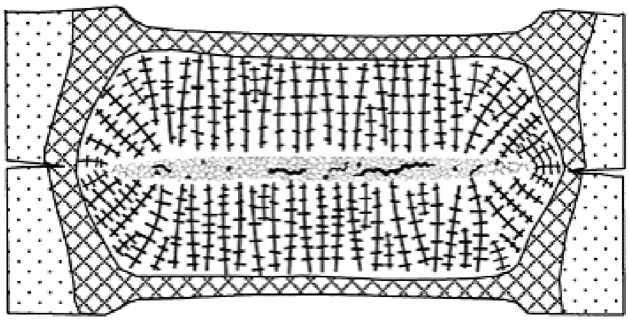


Figure 50- Structure obtained with overcooled electrodes condition [25]

The third case is verified when the cooling rate is faster in longitudinal direction rather than in vertical one. This can happen if the contact between electrodes and metal sheets is located in a small area or electrodes are subject to wear. The result is shown in figure 51. It is clear that the last liquid volume with equiaxed grains solidifies at the center of the nugget, since the most heat is dissipated through the sheets. In this area cracks and voids are located mostly, due to the reduced dimension of the mentioned volume, as before. Being far from the periphery of the spot weld, the effect on its strength, is neglected. However, the propagation originated from these discontinuities can become a problem if they arrive close to the edge of the nugget. [25]

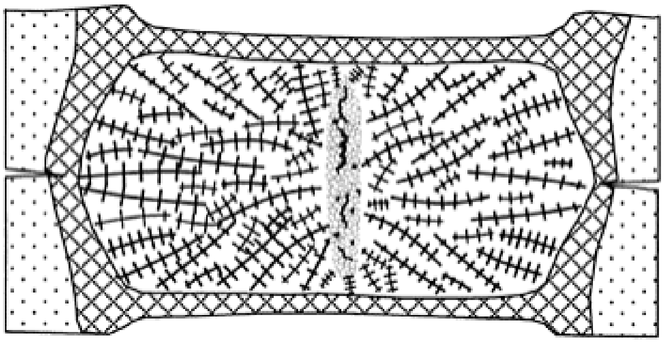


Figure 51-structure of the nugget with cooling rate faster along longitudinal direction [25]

4.4 Electrode wear

The most important point to focus on is the phenomenon of electrode degradation during spot welding. As the application field of the platform is predictive maintenance, when it comes to spot welding, the electrodes are the most critical components of the system as they are most subject to wear. In the following section, we will therefore describe what is meant by the term "electrode life" and which parameters have the greatest influence on this phenomenon.

4.4.1 Electrode pitting

Electrode pitting is a phenomenon responsible of degradation of electrode. It involves the tip face of the electrode during RSW: when the electrodes are removed from the sheets surface, at the end of the process, parts of electrode material fall off from the tip face. The consequent wear, affects what is defined as the life of the electrode, which is "*the first weld number at which the joint strength dropped below 80 pct of its initial value*" ([30]). For example, during a common welding process of aluminum alloy, the presence of oxide layer on sheet surface allows few points of contact between electrodes and sheets, forcing all the welding current to flow through them (when electrodes are put on sheets, fractures on oxide layer are produced, allowing contact for welding). However, forcing high current to flow in small points induces generation of excessive heating, that causes local melting of sheets and electrode materials together. The creation of copper-aluminum alloy is responsible of pitting. A deeper discussion on generation of pitting will be treated in the following of this section, supported by an experiment.

From the description above, it is possible to guess that the pitting phenomenon involves many steps as aluminum pickup, melting process between sheets and electrode materials and cavitation, which lead to reduction on electrode life and a consequent incomplete welding. Moreover, their influence on weld joints affects shear strength and are responsible of a non-linear trend of it. Indeed, due to electrode degradation, four stages are observed, depending on contact area changing. In the first stage, the contact area is constant and so the shear strength of the weld joint. In the second stage, the strength starts to increase, initially due to the aluminum pickup and then as the consequence of the increase of contact area. At the end of this stage, the strength reaches its peak and pitting occurs. The growth of already present pitted areas in the previous stage and the birth of new in the third stage, involves a further increase of the contact area and the generation of cavitation, phenomena that occurs when more pitted areas groups together. The result is a rapidly reduction of welded joints strength. In the last stage, the produced joints are incomplete and its strengths are strictly dependent on morphology of pitted area, now varies and unpredictable. The electrode has reached the end of life.

The stages described above can be easily recognized in the following study, conducted in order to observe electrode degradation during welding of aluminum alloy. The experiment focuses on the top electrode, since it is subject to faster wear. Its tip face morphology changes quicker due to the heat generation. Indeed, being the top electrode the positive pole and the bottom electrode the negative, heat generation affects most the upper one during welding, and its life results shorter. The electrodes chosen for the study, with diameter tip of 10 mm, radius of curvature of 50 mm and taper angle of 60 deg, is shown in figure 52. The specimens used have thickness of 1.5 mm of aluminum alloy AA5182-H111. 10 percent of welded joints up to 500 are periodically tested to evaluate shear strength, 5 percent after 500 welds. Then, the same experiment is repeated twice, but with a cleaning operation every 20 welds and 50 welds, in order to evaluate any improvement on electrode degradation, induced by the removal of aluminum alloy pickup on electrode tip face.

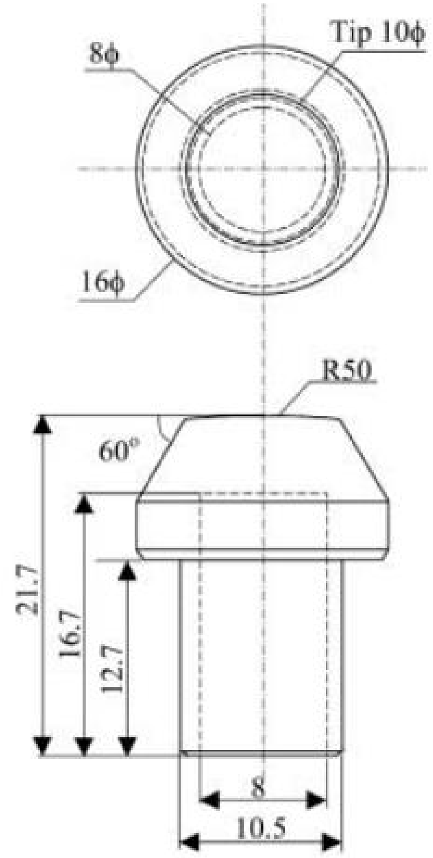


Figure 52-Sizes in mm of electrode used [30]

The surface of the same electrode tip is viewable in figure 53, where a pic of the face is taken after 20 (a), 50 (b), 100 (c), 200 (d) and 500 (e) welds. The contact area od welding is clearly visible on tip face, since it turns to gray due to aluminum alloy pickup. Looking at figure 53 (d), it evidences that pitting starts at the periphery of contact area until to form a ring around it. When the ring is completely formed, pitting begins to expand toward the center on the contact area causing cavity figure 53 (e) and 53 (f). The same phenomenon can be recognized in figure 53, where carbon imprints are shown.

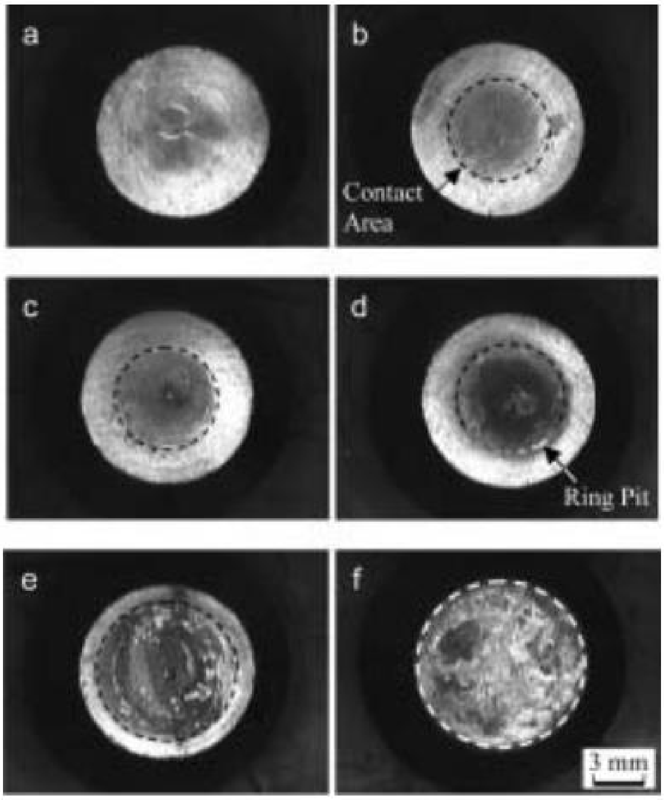


Figure 53 – Electrode face tip after 20 (a), 50 (b), 100 (c), 200 (d) and 500 (e) welds [30]

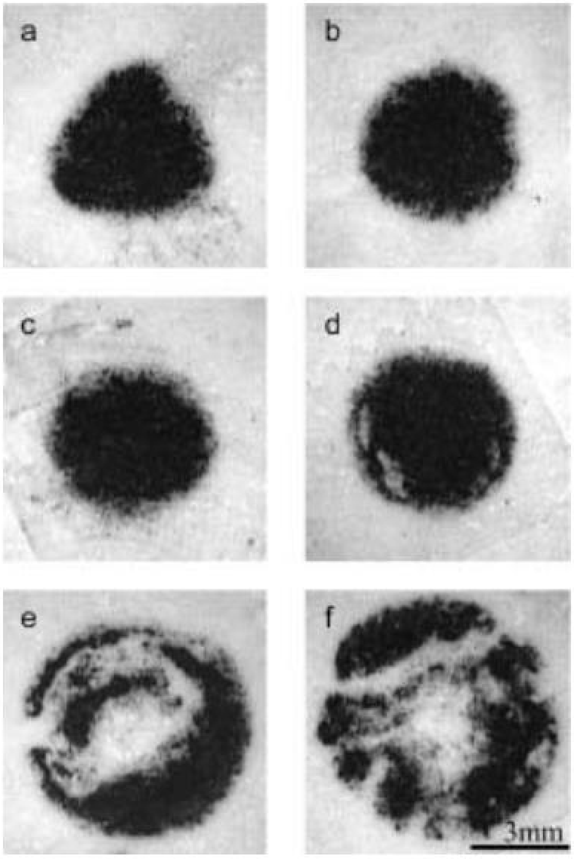


Figure 54 - Carbon imprints of electrode after 20 (a), 50 (b), 100 (c), 200 (d) and 500 (e) welds [30]

Through the EDX (Energy Dispersive X-ray Spectroscopy) analysis, a participle of the same material composition of material sheet is found out on the electrode face tip from the first weld already, that means the pickup phenomenon starts at the beginning of the welding process. Due to the current constriction between sheets and electrodes, the heat generated is enough to melt and weld together the aluminum pickup with the electrode face tip. After 200 welds, welding of aluminum with copper is responsible of formation of different layers on electrode face tip with different composition. Thanks to deeper analysis, material compositions and chemical characteristics are detected. Different layers have different material composition due to different phase at which Cu-Al alloys are weld together. The only easily identifiable layer is $CuAl_2$ thanks to its thickness. Due to very short weld time, the amounts of intermetallic phases have very thin layers, since they may be located only in local spots and they need more welding time to melt in a proper way. So, they result too small to be clearly identified. Anyway, local spots of Cu_9Al_4 are more frequently detected.

It is possible to believe that pickup and alloying phenomenon described above is responsible of the initial strength's increase. The pitting can be verified in two manners: the first one is shown in figure 55, where intermetallic phase bonds break in local regions, causing material removal from electrode face tip. Looking at the sheet surface, the presence of particles with the same shape of the pitting confirms what has been mentioned; the second one is the result of electrode material melting. During welding, material on electrode face tip binds with the aluminum on the sheet surface and remains onto it when electrodes are separated from the sheets. Figure 56 shows the electrode pitting and electrode material transferred onto the sheet. from the image it is evident how the two shapes match perfectly.

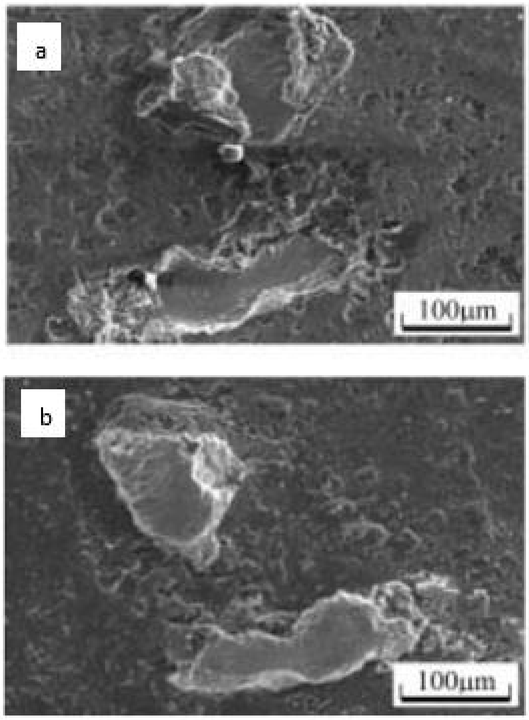


Figure 55 – First pitting type, electrode surface with two holes (a) and sheet surface with two foreign particles (b) [30]

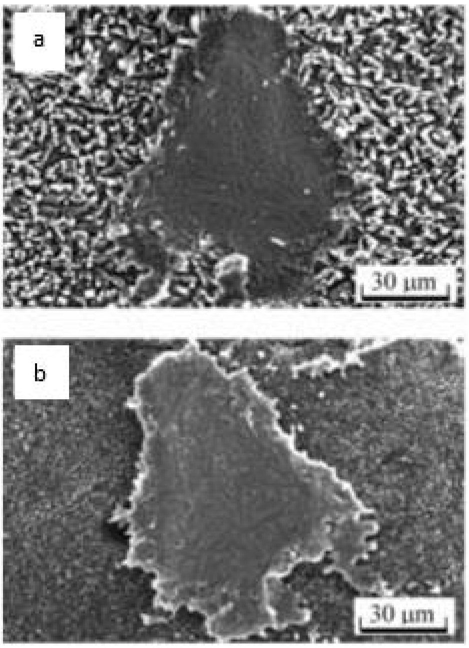


Figure 56- Second pitting type, electrode surface with a hole (a) and sheet surface with a foreign particle (b) [30]

When holes are formed due to pitting, they start to growth with subsequential welds. In this experiment, three holes are located on the electrode face tip after 76 welds, but reaching 85 welds, the three islands are so large that they come together to form a single larger. Then, cavitation phenomenon takes place after 85 welds. Figure 57 shows this situation.

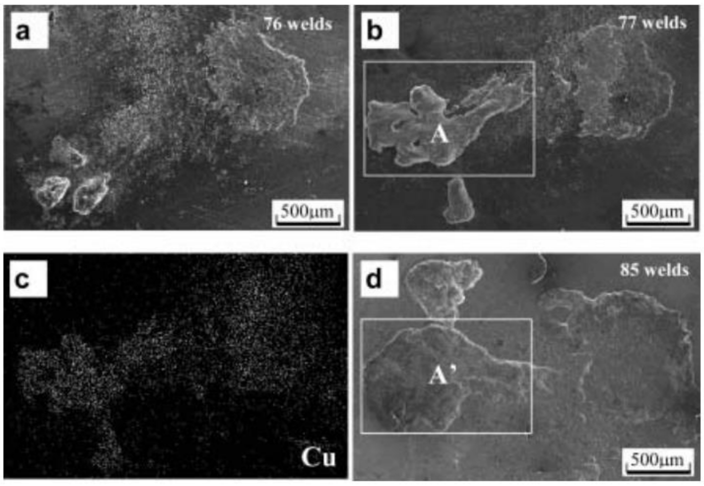


Figure 57 – Cavitation in RSW. Holes at 76 welds (a), holes' growth at 76 welds (b), mapping of Cu (c), cavitation at 85 welds (d) [30]

Therefore, pitting on electrode face tip can be explained in the following way: the total amount of heat generated by welding current in few points of contact, induces the melting of aluminum alloy and transfer it onto the surface of face tip. These particles react with electrode material forming layers of Cu-Al mixture. When the weld joint is done, electrodes are separated from sheets and part of the electrode material is removed from the face tip. This can happen due to the breaking of intermetallic alloy bonds, if the mixture is solidified before the end of welding process, or due to the transfer of molten alloy that remain on sheet surface before the separation of the electrodes from sheets.

The last operation of this study is to analyze electrode degradation on electrode cleaning the face tip every 20 or 50 welds. The improvement on performance of the electrode is easily guessed, but a deeper analysis shows that a good compromise is cleaning the electrode every 30-40 welds. Figure 58 shows the difference between the electrodes in all these situations at the end of welding process in comparison with a new electrode. Figure 59 illustrates the corresponding carbon imprints.

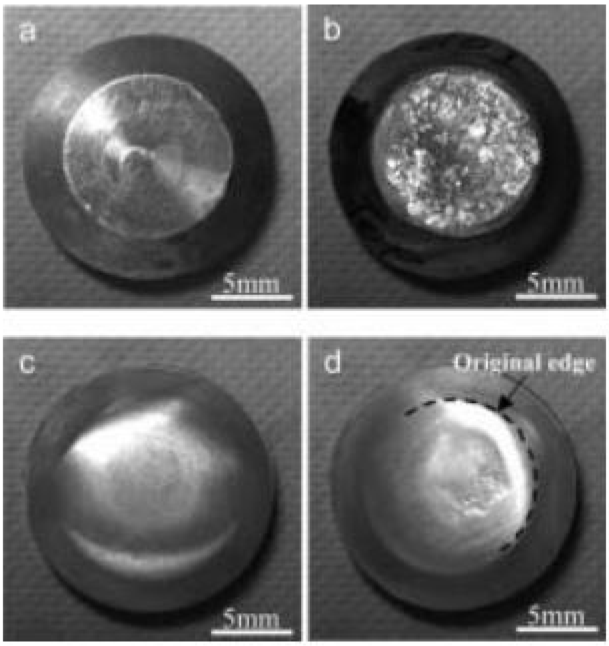
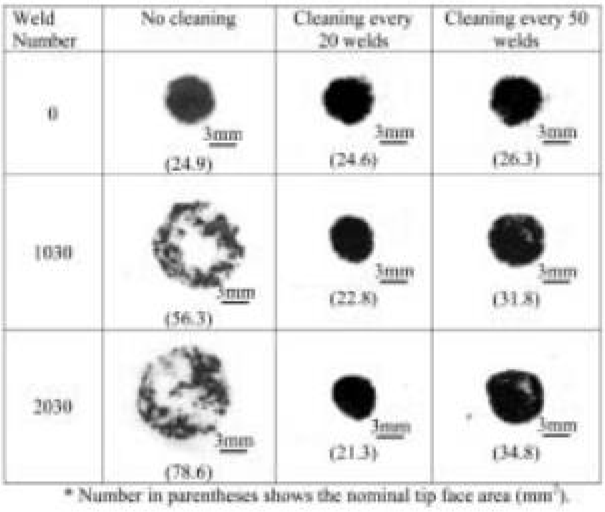


Figure 58 – Surface of new electrode (a), surface of electrode after 2000 welds without cleaning, with cleaning every 20 welds (c) and with cleaning every 50 welds (d) [30]



Weld Number	No cleaning	Cleaning every 20 welds	Cleaning every 50 welds
0	3mm (24.9)	3mm (24.6)	3mm (26.3)
1030	3mm (56.3)	3mm (22.8)	3mm (31.8)
2030	3mm (78.6)	3mm (21.3)	3mm (34.8)

* Number in parentheses shows the nominal tip face area (mm^2).

Figure 59 – Carbon imprints of electrode without cleaning and with cleaning every 20 or 50 welds [30]

As it can clearly see, after 2030 welds, the not cleaned electrode is completely damage, but the initial edge is still visible. The electrode with cleaning every 50 welds, is less damage rather than the previous one, due to the ability to delay the increase of the contact area, and keep lower the current density, allowing nugget formation properly. The result of cleaning every 20 welds is unexpected: the pitting is absent even after 1000 weld (value at which it takes place with electrode cleaned every 50 weld - for not cleaned electrode, pitting is located at 360 welds yet-), but the quality of weld joints is badly affected after

930 welds due to some detected alteration on tip geometry. This alteration appears as distorted electrode profile (the edge of surface tip is lost in part) and has negative impact on nugget formation. [30]

4.4.2 Pitting on uncoated steel vs DP600 steel

Although pitting is a phenomenon affecting all electrodes, it does not affect their lives to the same extent. With the following study, it will be checked how the pitting is responsible for the degradation of the electrode according to which material will be welded. In this section, electrode wear is analyzed, welding 0.8 mm thickness of galvanization dual-phase DP600 and using similar specimens of uncoated steel. The study focuses on radial and axial wear, comparing the obtained results under the two conditions.

For the experiment, sizes and shape of spherical electrodes used are shown in figure 60, while welding parameters are setting in according to table 1. Welds are produced keeping a distance of 20 mm each other and are subject to tensile shear tests to evaluate how electrode degradation affects the quality of welded joints. Only 50 welds are tested until 300 welds, then 100 welds once this value is exceeded. Carbon imprints of electrode face tip are taken to investigate the occurrence of the pitting phenomenon, instead radial and axial wear are obtained through a measurement sensor with precision of 0.01 μm, mounted on the AC servo welding gun used during the weld process.

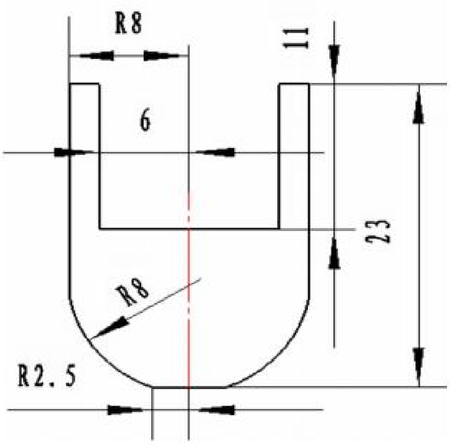
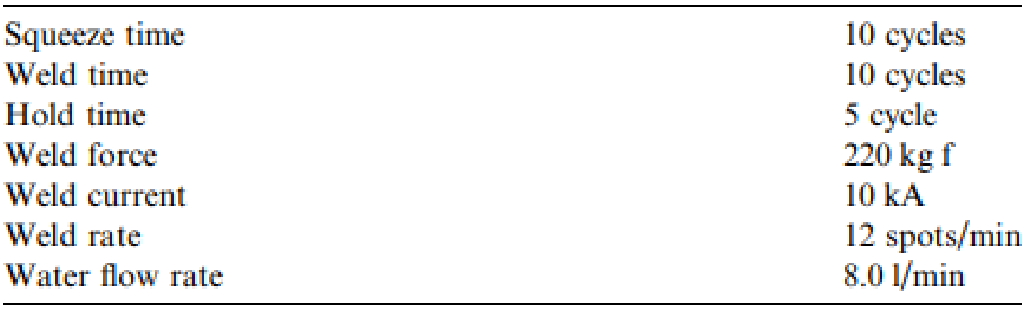


Figure 60 – Sizes and shape of electrodes (in mm) [31]



Squeeze time	10 cycles
Weld time	10 cycles
Hold time	5 cycle
Weld force	220 kg f
Weld current	10 kA
Weld rate	12 spots/min
Water flow rate	8.0 l/min

Table 1- Welding parameter for DP600 of AC servo welding gun [31]

The results of the study highlight that welding DP600 steels reduces electrode life, accelerating electrode wear. This behavior is guessed due to the need of this steel of more time and current to weld. Looking at carbon imprints of figure 61, it can be seen that pitting occurs at about 200 welds (white regions on electrode face tip). Thanks to carbon imprints the contact area is clearly visible and an enlargement of it is detected (which corresponds with an enlargement of electrode face tip diameter) with the increase of weld spots.

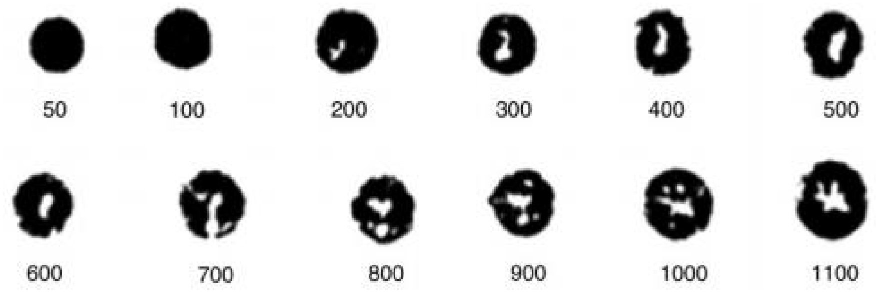


Figure 61- Carbon imprints of electrode face tip welding DP600 steel [31]

Graph in figure 62 plots radial wear for DP600 steel and uncoated low carbon steel in function on the number of welded joints.

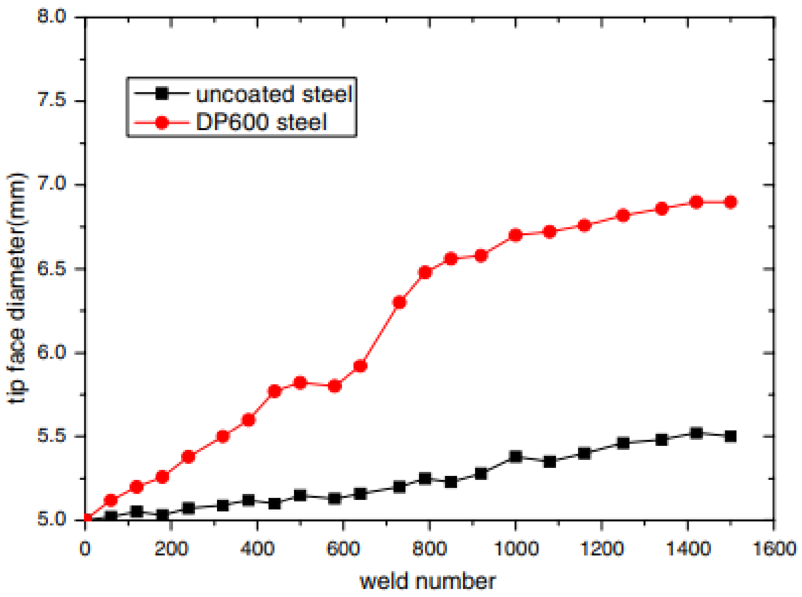


Figure 62-Tip face diameter on number of welded joints [31]

As predicted, radial wear of electrode used for welding uncoated steels is less than DP600 steels. Moreover, in DP600 steels, a great step is detected around 700 welds, probably due to cavitation: when pitting areas grouped together, cavity takes place and under high thermal conditions, it breaks off due to high pressure from electrode force. The result is a sudden enlargement of contact area (and consequentially increase of diameter tip), now able to withstand the previous electrode force. This event also induces a decrease of current density, that leads to a reduction of radial wear rate after 700 welds (1.85 µm for weld after 700 welds in comparison with 0.62 µm after that value). The total enlargement of diameter tip is about 1.9 mm (from initial value of 5 mm to final value of 6.9 mm) during whole electrode life, ended after around 1500 welds. Looking at the diameter tip trend for uncoated steels, it is possible to conclude that radial wear is slower than the DP600 steels: electrode life ends around 8000 welds, radial wear rate is about 0.35 µm until 1500 welds, and the total enlargement of diameter tip is about 0.52 mm (from initial value of 5 mm to final value of 5.52 mm).

Similar results are obtained analyzing the behavior of axial wear.

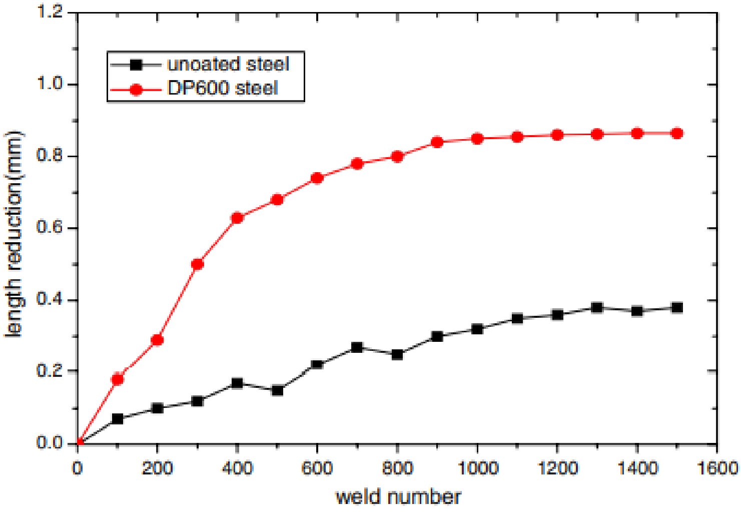


Figure 63 - Axial wear of electrodes welding DP600 or uncoated steels [31]

Axial wear rate is still faster for DP600 steels and that induces shorter electrode life. As illustrated in figure 63, the axial wear rate has two possible values: before 400 welds it is about 0.78 µm, after that, it decreases to 0.23 µm. As before, this change in rate is due to more copper removal from electrodes as a consequence of high current density and heat generated in the first part of the process with respect to the second one. The total length reduction is about 0.86 mm taking into account the whole electrode life. Values of uncoated steels are all smaller: axial wear rate is equal to 0.25 µm until 1500 welds and the electrode life can reach 8000 welds. The total length recorded down to 1500 welds is equal to 0.38 mm.

Since the quality of welded joints is most influenced on electrode diameter enlargement rather than other parameters, radial and axial wear rates can be useful to estimate and predict the growth of electrode diameter face tip. Keeping in mind that the electrodes used in this experiment are spherical, the following formula can be applied (figure 64, at the end of the section, clarifies how this expression is obtained):

$$r_1^2 = R^2 - \left(\sqrt{R^2 - r_0^2} - \Delta h\right)^2 \quad (2)$$

where $r_1$is the new radius of face tip after welding, r_0 is the initial radius of face tip, R is the spherical radius of the electrode, given by electrode's shape and Δh is the total length reduction.

As said before, the effect of electrode wear causes a reduction on welded joints strength. In particular, for DP600 steels, tests reveal that a good nugget is obtained at 200 welds, with 9.3 kN strength, indentation of 360 µm of depth and diameter of 5.8 mm. However, at 1200 welds, the joint is incomplete, due to a not proper penetration (indentation of 90 µm only) and reduction strength to 8.2 kN.

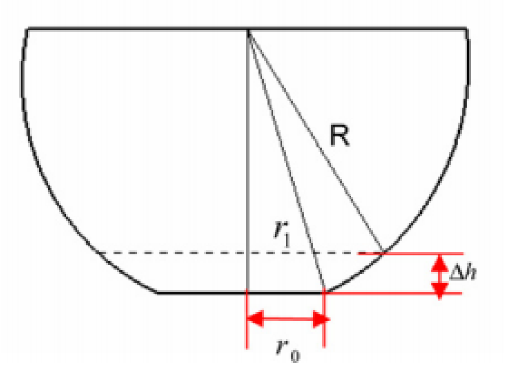


Figure 64-Electrode spherical shape [31]

Chapter 5: Laboratory experience

After the design of the platform, a section dedicated to the description of the laboratory experience follows. This chapter describes the equipment that will make up the system on which the platform will operate and the methodology used to collect the data to be computed, in order to test a future implementation of the platform. As this platform is intended to predict the degradation of electrodes in a resistance spot welding machine, the main focus was on studying them, taking into account all the studies carried out in recent years presented in the previous chapter. This had a strong influence above all on the choice of materials used and the working conditions, aimed at recreating as far as possible a situation of machine activity similar to the industrial one.

5.1 Collection of incoming data

In order to collect data to test future operation of the platform, a large number of joints were produced to wear out a pair of 0.6 mm diameter copper electrodes mounted on an industrial resistance spot welding machine. A "Tecna Medium Frequency Welder" was used as machine for welding, equipped with sensors grouped in table 2. Particular attention has to be focus to the panel PC weld monitor, as it is equipped with an internal processor capable of carrying out an initial processing of the data coming out of the machine. It is thanks to this PC that the machine can be integrated into an IoT system, as it has an IP address that allows it to be connected and exchange its data with the platform and the other smart devices in the system. The machine is therefore equipped with multiple controls, such as position sensor, pressure sensor, etc. aimed at acquiring the data produced during the state of activity, while the use of the PC is not only to process the data but also to allow an initial display to the operator. There are many data involved, the most important being those that most influence the quality of the joint discussed in the previous chapter, i.e. the current trend, the electrodes force, on which the indentation depends, and so on. A representation of some of them is shown in figure 65, similar as they are presented to the operator via the panel.

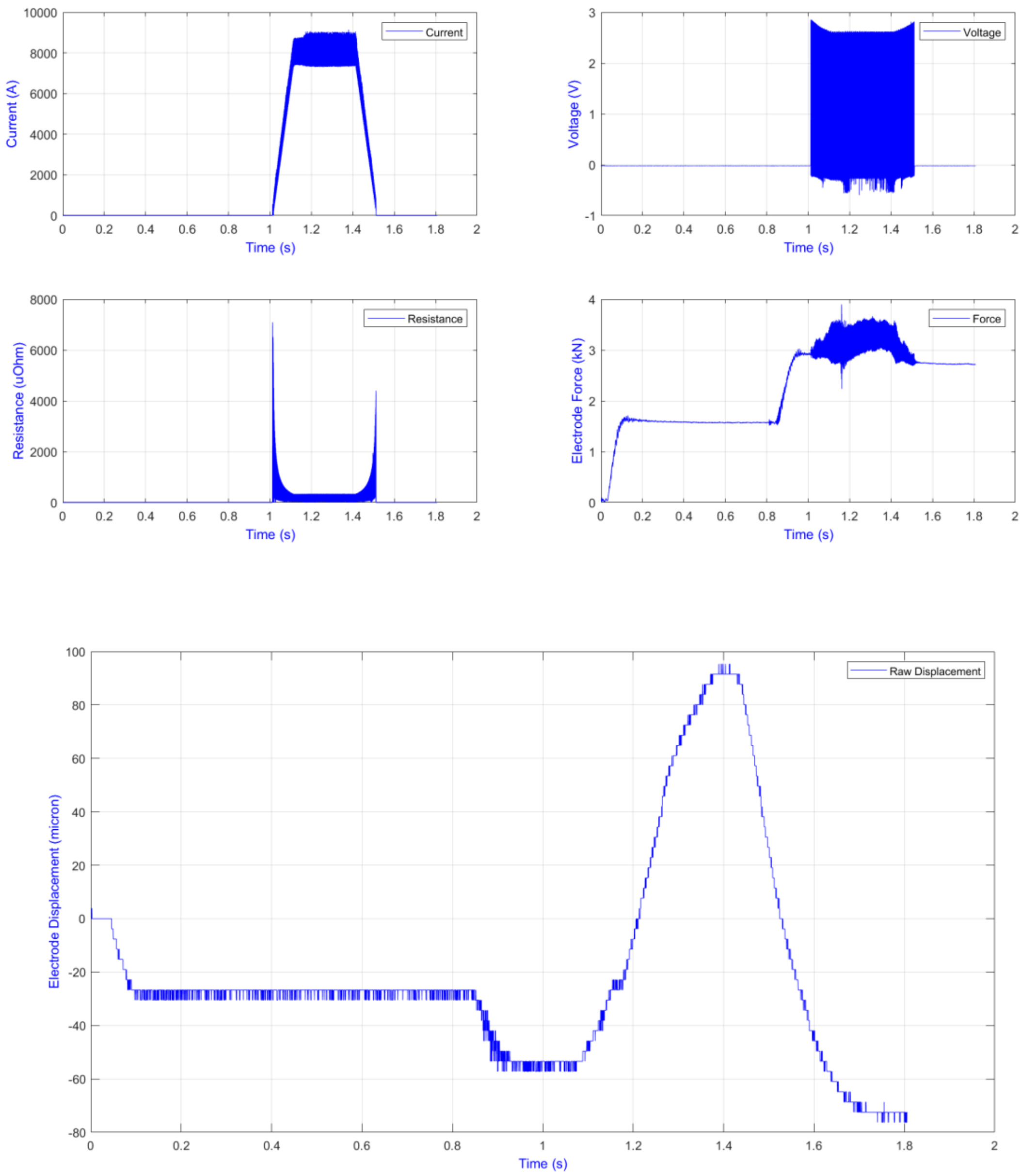


Figure 65 – Trends of welding process signals over time

For example, the last graph in the image above was extrapolated due to the presence of the position sensor. This sensor, which is accurate to the nearest micron, is shown in the following image and the result of its measurement provides the displacement of the electrode during welding.

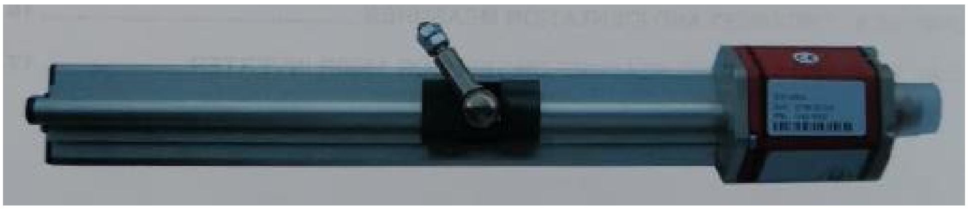


Figure 66 - Linear position sensor item 23480, with resolution of 2 µm

The model of the machine is shown in the image 67, while its datasheet is not included in the discussion as it is easily accessible.

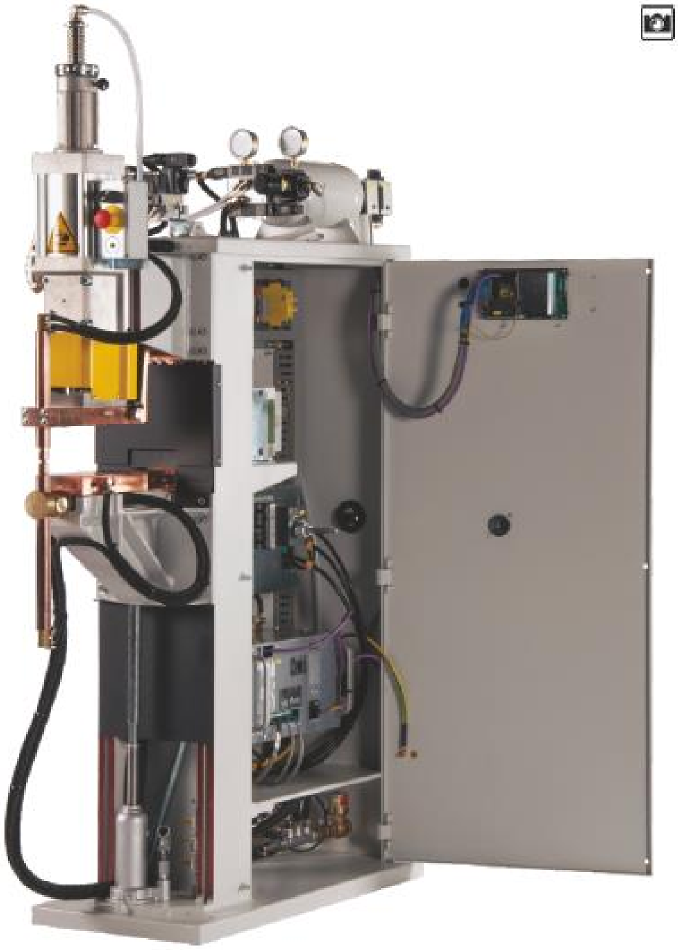
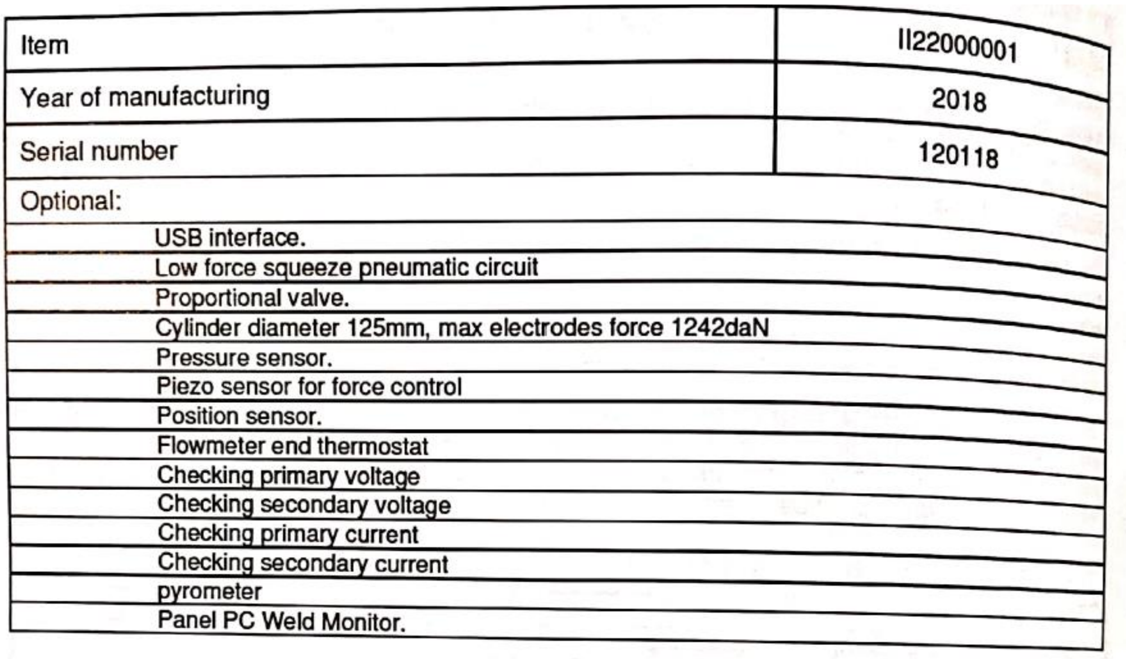


Figure 67 – Tecna Medium Frequency Welder item II22000001



Item	II22000001
Year of manufacturing	2018
Serial number	120118
Optional:	
USB interface.	
Low force squeeze pneumatic circuit	
Proportional valve.	
Cylinder diameter 125mm, max electrodes force 1242daN	
Pressure sensor.	
Piezo sensor for force control	
Position sensor.	
Flowmeter end thermostat	
Checking primary voltage	
Checking secondary voltage	
Checking primary current	
Checking secondary current	
pyrometer	
Panel PC Weld Monitor.	

Table 2 – Sensors equipped with the resistance spot welding machine

There is also a pyrometer associated with the machine, which is used to acquire information about the temperature trend during welding around the welded spot. Once acquired, this data is also forwarded to the PC for initial processing. Further temperature monitoring is performed by a “FLIR SYSTEM” thermal imaging camera (Figure 68), specially allocated to capture the temperature variation during the process. It is equipped too with a PC and an IP address so that it can be integrated into the IoT system. The image captured by the camera is displayed on the associated PC and is shown in the figure 69.

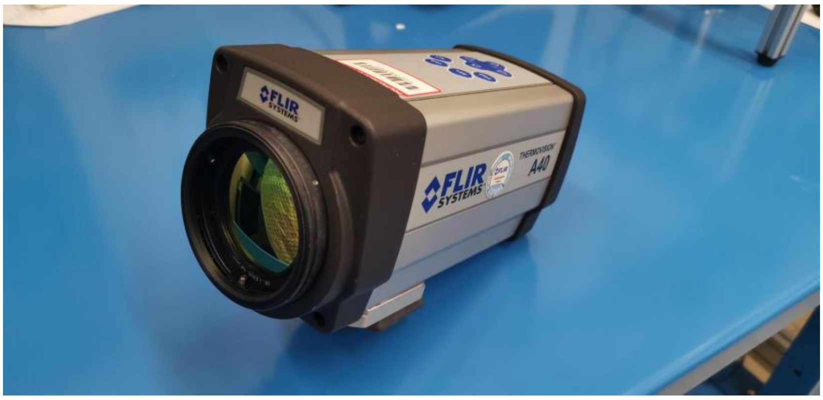


Figure 68 – Thermal imaging camera - Termovision A40M

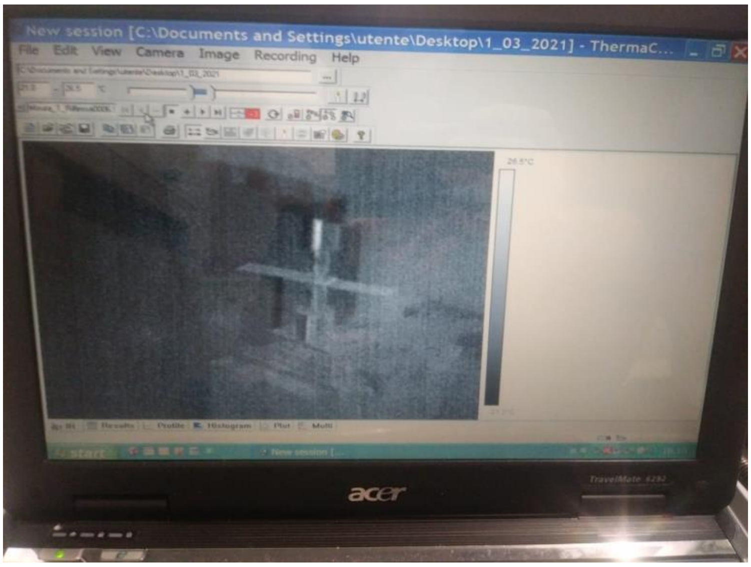

New session [C:\Documents and Settings\utente\Desktop\1_03_2021] - ThermaC...
File Edit View Camera Image Recording Help
start
acer
TravelMate

Figure 69 - PC display associated with the camera

Image 70 shows the situation recreated in the laboratory in which the joints were made. It is possible to recognize the welder in the center, the pyrometer on the left and the thermal imaging camera on the right, in the foreground.

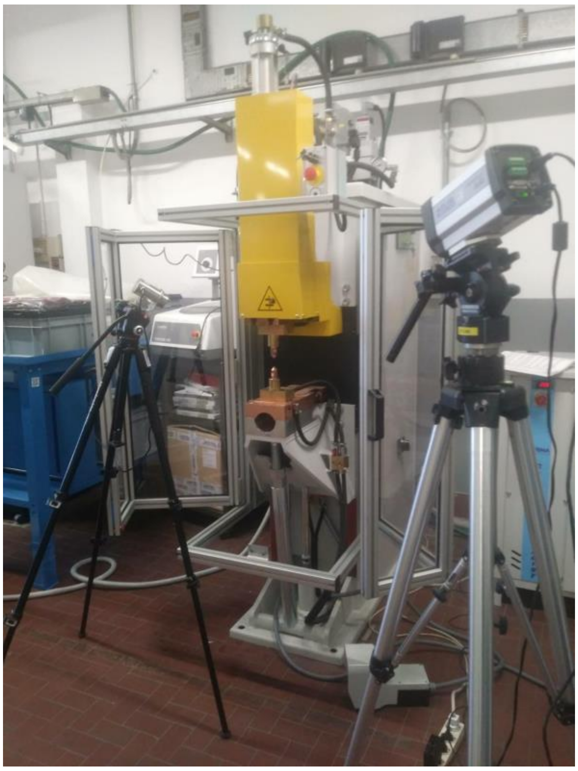


Figure 70 – Laboratory conditions in which the welds were carried out

DP600 was chosen as the material for welding the joints because it is the most widely used in the manufacturing industry for its strength and lightness. For the preparation of the sheets of 0.8 mm of thickness, the AWD D8.9M standard was followed, which provides a total of 135 joints per sheet distributed over 9 rows of 15 points each. The dimensions of the joints and their geometric distribution on the sheet is always imposed by the standard (Figure 71). The standard also provides information on the sequence in which to weld the joints, since a different order could cause distortions on the latter that would alter the properties of the material, thus invalidating the results of the study.

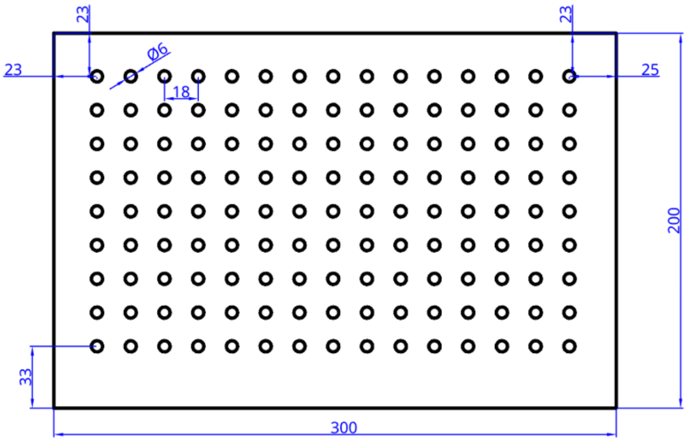


Figure 71 - dimensional indications of the sheets provided by the standard

The following image shows a sheet of DP600 metal after all 135 welds have been completed. The machine parameters used to make the joints were set in line with those used by industrial companies, in order to simulate a real application of the machine. In particular, the following values was chosen:

- current rise time: 100 ms
- welding time at full current: 300 ms
- current descent time: 100 ms
- steady-state current intensity: 8 kA
- electrode pressure: 1.5 bar (equal to an electrode force of 3.5 kN)

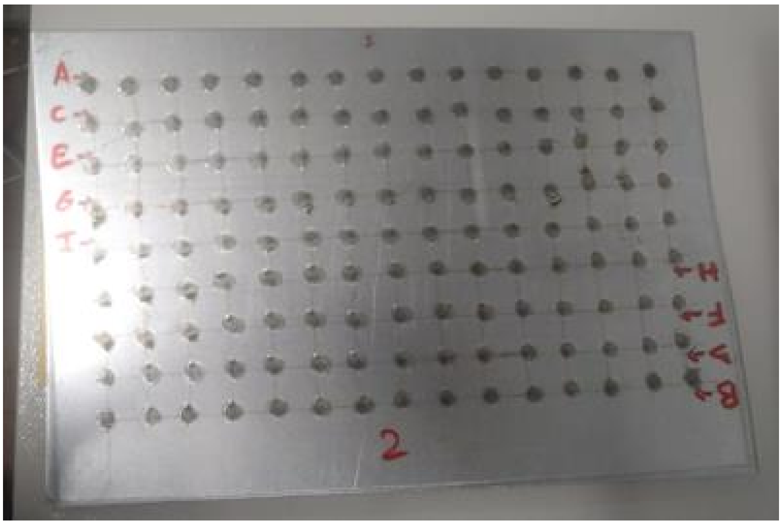


Figure 72 – DP600 steel after welding

In order to check when the degradation of the electrodes was so advanced as to render the electrode unusable, three joints were made every time a sheet was completed, i.e. every 135 joints. The dimensions of the specimens are shown in the following image, and its sizing was specifically chosen to meet the standards dictated by the regulations. Specifically, they are 10 cm long and 3 cm wide. The overlap of the two sheets is marked by a black line. Figure 74 shows the specimens just before being welded.

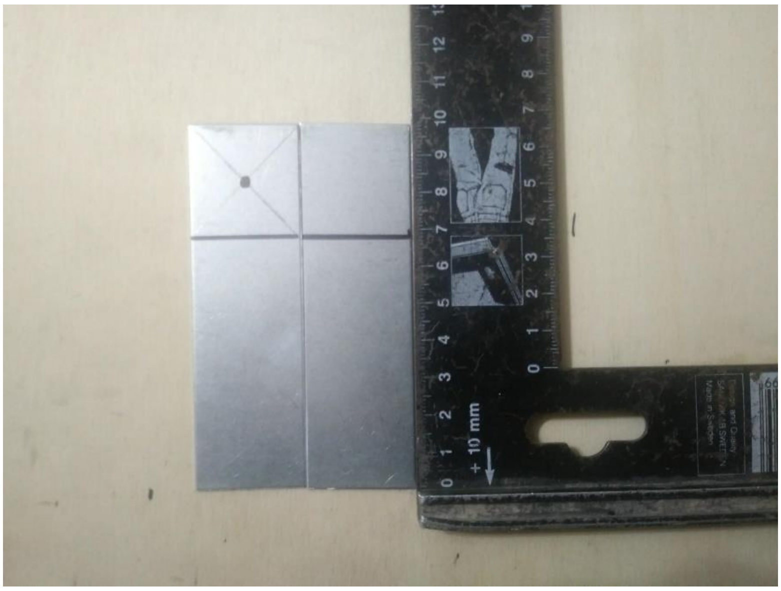


Figure 73 – Specimens of DP600 for degradation evaluation

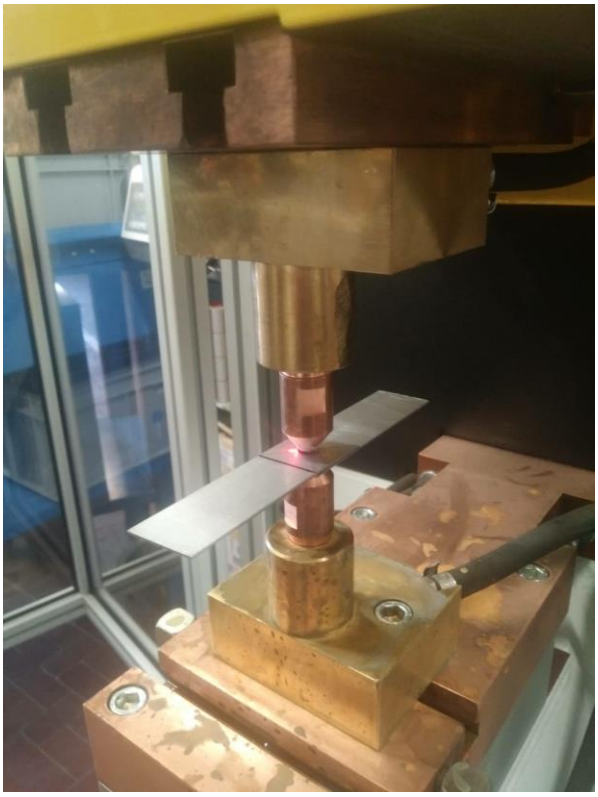


Figure 74 - Specimens under spot welding

Once welded, the specimens were subjected to a tensile test to failure using a special machine capable of measuring the force required to break it. This parameter can be used to determine when the electrodes have reached the end of their life, as they are no longer able to produce joints with the strength characteristics required by the manufacturers. The discussed machine is an “EASYDUR” 3MZ model (Figure 75).

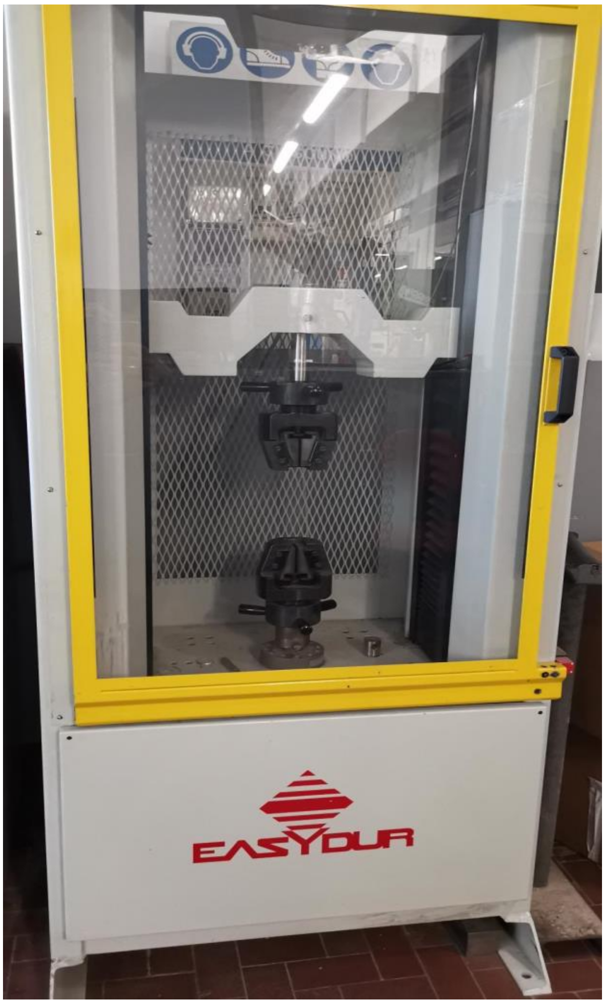


Figure 75 - Testing machine -EASYDUR 3MZ model

In addition, every 15 welded joints, i.e. every time a row of joints was finished, the impression of the electrodes was collected through carbon paper, to check the change in size and geometric shape of the electrode tip face due to wear. Through carbon paper it is also easy to see when pitting occurs. It is in fact necessary to observe at image 76, which shows this phenomenon at the 1000th welding point, but its first appearance occurred on the 200th one.

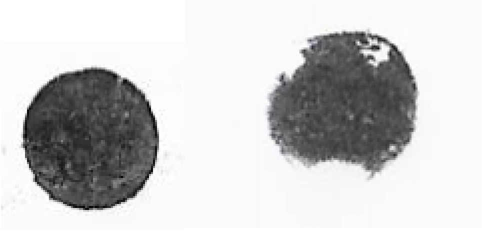


Figure 76 - carbon imprints of electrode tip face at 0 welded joint (on the left) and at 1000th one (on the right)

The pitting phenomenon is also easily recognisable by observing the tip face of the electrodes with the naked eye. Figure 77 is shown to highlight the change that has taken place at a distance of just two hundred points. The first picture refers to the condition of the electrodes at around ten welded points, while the second picture refers to around two hundred and fifty welded joints. The pitting phenomenon is evidenced by a lighter colour, tending to white, in the affected area on the electrode tip face.

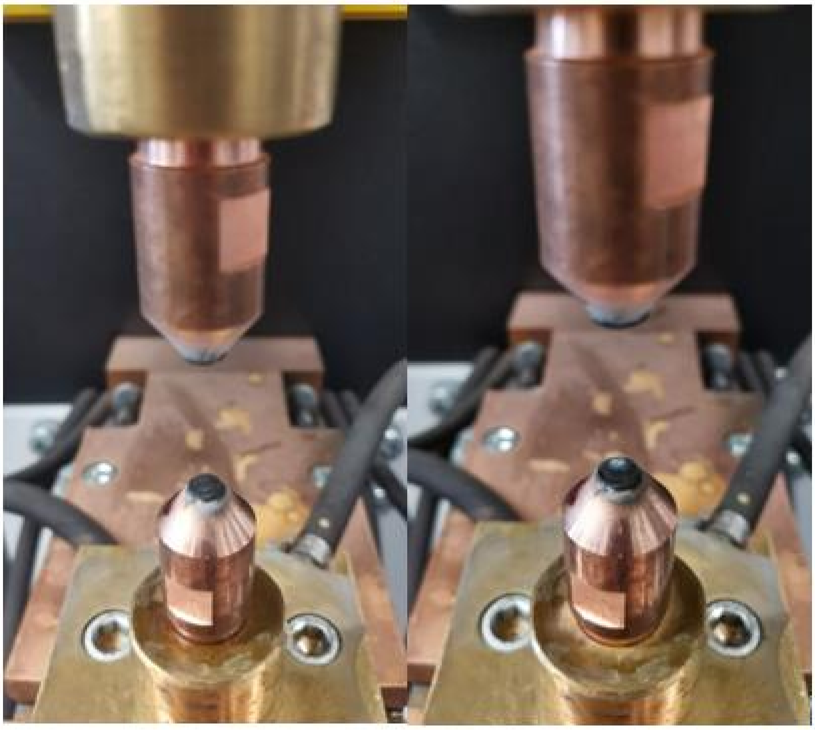


Figure 77 - Electrode tip face at 10th welded joint (on the left) and at 200th one (on the right)

The study showed that the electrodes, under the given conditions of activity set out above, reached complete wear around the 800th joint.

5.2 Platform application on case study

Now that the physical components that make up the plant have been accurately presented and the data of interest in the case study have been identified, the future operation of the designed platform is presented in broad terms. By adapting the design of the platform to the system defined above, all the virtual components that reside in the edge layer will have to be implemented as close as possible to the real components in order to exploit edge computing technology. There are therefore two possibilities:

- Exploiting the welder's PC as a computational entity: being to all intents and purposes a PC, it is able to host the components of the edge layer. In this case, since it is equipped with an internal processor, it will be possible to apply the model for predictive analysis coming from the "*Predictive Analytic Service*" hosted by the cloud and process the data coming from the machine at run-time to predict the degradation of the electrodes.
- Exploit the presence of the switch to which all system devices connect: in this case, the switch would have two functions. The first is to receive all the information packets coming from all the devices belonging to the physical system (typical functionality of all routers) and to address them to another network, the one to which the cloud belongs; the second is to carry out the computational and data processing activities as described in the previous point for the welder's PC.

From a design point of view, the two solutions are very similar, although theoretically the first solution would be preferable as the data transfer would be reduced. However, this is a choice to be made during the platform implementation phase, as there are many parameters to be taken into account, such as the computational capacity of the computer's processor, the available memory and so on. If the first solution is chosen, the camera PC must be allowed to send the data to the welder's PC, so that the welder's PC can also take into account the results of the camera when applying the model from the "*Predictive Analytic Service*". This data exchange is possible because both computers have an IP address belonging to the same subnet. As for the machine used to test the specimens, it will not be part of the system for which the platform was designed, as its purpose is only to return the true end of life of the electrodes, which will have to be predicted by the IoT system. Its operation is therefore limited to the laboratory experiment itself, in order to have a feedback of the correct prediction of the IoT platform.

Even if the first option is chosen, the presence of the switch is still essential, as in addition to allowing communication between devices belonging to the same sub-network, it also allows the exchange of information between the devices and the cloud, which also has an IP address and is hosted on a server owned by the cloud providers. As described in section 3.2, it is in the cloud that data is processed in more detail and stored permanently. In fact, the database in the cloud has to host all the historical data coming out of the devices in the system since the new electrodes were mounted on the machine. Therefore, unlike those in the edge layer, which only store data temporarily and can therefore be very small, the cloud data storage must have sufficient space to store the data for all of them extrapolated by at least a thousand of welded joints. These data will then be used by the services in the cloud as described in detail. A note has to be made for the "I*mage Processing Service*". Its use is easy in this application as the electrode prints taken through the carbon paper has to be digitised and integrated into the system using a strategy that does not involve the use of a PC. As this data is collected manually by the operator on line on a sheet of paper, it requires an extra step to be integrated. Therefore, the operator must be equipped with a device capable of capturing an image of the electrode tip face on the sheet (e.g. using a smartphone) and uploading it to the dashboard on the machine, which has been developed as a web application and therefore also has an IP address. Thus, the dashboard will be used in two ways:

- has to be communicate with the user, allowing the display of the electrode degradation progress and alerting with appropriate alarms when the replacement time comes;
- will allow the user to upload images of the electrode imprint. Here, the images will be forwarded to the "I*mage Processing Service*", which, thanks to the application of special algorithms by the processor hosted in the cloud, will make it possible to extrapolate data such as eccentricity, diameter, pitting size, etc., which will then be stored in the database for future processing by the "*Predictive Analytic Service*".

If deemed necessary, images uploaded by the user could also be stored by adding a dedicated repository into the cloud.

Conclusion

The platform has been designed with the aim of providing the user with information regarding the wear of one or more system components and to support him in making decisions arising from it, regarding the planning management of tasks. It can be easily extended and adapted to a multitude of production processes. What characterises one application from the others are the algorithms that make up the services, which must be designed ad hoc for each of them. This has not been discussed here as it does not concern the designing phase of the project, but more properly the implementation phase, for which more advanced information technology (IT) skills would be required.

In this particular case, the IoT system aims to provide an estimate of electrode wear in a resistance spot welding machine. The model used to compute the data at run-time has to return to the user, through the dashboard display, a graph similar to the one shown in figure 78, where the y-axis shows the reduction in the resistant force of the joint in percentage terms (reference term of the electrode life) and the x-axis shows the number of welded joints. In red is indicated the trend of the predicted wear, from the zero point, the joint from which the model was applied for the first time (instant in which the prediction started) up to the horizon of invalidation of the electrode. In black, the actual values obtained from run-time data computation are recorded. The joint strength parameter is not directly derived from a measurement but can be obtained indirectly from all other parameters, through the algorithms mentioned above. Should the system detect a progressive deviation of the real trend from the predicted one, the cloud would provide the edge layer with a model on which to base the prediction that is more up-to-date and suitable for the operating conditions of the machine, obtained thanks to the processing of the data used to extrapolate the old model with the addition of the new data acquired from the last welds.

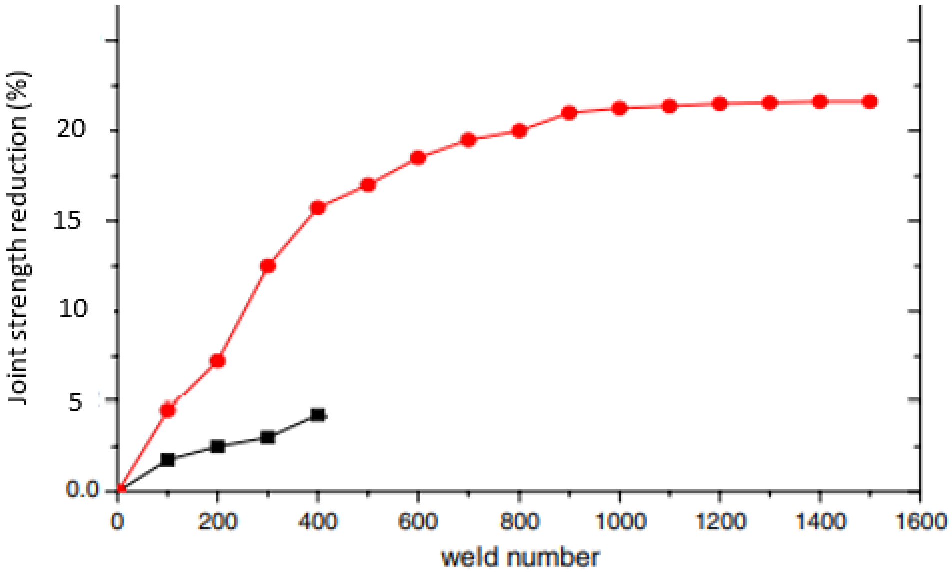


Figure 78 -Example of degradation electrode's graph

The platform is still in the "predictive maintenance" paradigm, since although it does not perform the function of a mere prediction of a degradation process (having the scheduling service) it is not able to solve in an autonomous way the problem of wear and tear, but requires the replacement of the electrodes with a new pair. Therefore, not being able to avoid wear and tear and not carrying out maintenance activities in an autonomous way, the IoT system can be considered, as a whole, a precursor of the "prescriptive maintenance" paradigm, even though it is not part of it.

References

[1] J. Lee, M. Azamfar, J. Singh, J. Feng, B. Jiang e J. Ni, «Intelligent Maintenance Systems and Predictive Manufacturing,» July 2020.

[2] J. Yoon, «Deep-learning approach to attack handling of IoT devices using IoT-enabled network services,» 2020.

[3] Y. Hajjaji, W. Boulila, I. Farah, I. Romdhani e H. A, «Big data and IoT-based applications in smart environments: A systematic review,» 2021.

[4] I. S. P. U. 2005, «ITU internet reports: The internet of things,» *Proceedings of the International Telecommunication Union (ITU),* 2012.

[5] A. Tamboli, «Build your Own IoT Platform,» 2019.

[6] A. Katal, M. Wazid e R. Goudar, «Big Data: issues, challengers, tools and good practices,» *Proceedings of the Sixth International Conference on Contemporary Computing ,* 2013.

[7] M. Fahmideh e D. Zowghi, «An exploration of IoT platform development,» 2020.

[8] M. Buvana, K. Loheswaran, K. Madhavi e A. Behura, «Improved Resource Management And Utilization Based On A Fog-Cloud Computing System With IoT Incorporated With Classifier Systems,» 2020.

[9] L. Beno, R. Pribis e R. Leskovsky, «Processing data from OPC UA server by using Edge and Cloud computing,» 2019.

[10] S. Bello, L. Oyedele, O. Akinade e M. Bilal, «Cloud computing in construction industry: Use cases, benefits and challenges,» 2021.

[11] Y. Mansouri e B. M, «A review of edge computing: Features and resource virtualization,» 2021.

[12] F. Bonomi, R. Milito, S. Addepalli e J. Zhu, «Fog computing and its role in the internet of thing,» *Proceedings of the 2012 ACM First Edition of the MCC Workshop in Mobile Cloud Computing,* 2012.

[13] H. Sabireen e V. Neelanarayanan, «A Review on Fog Computing: Architecture, Fog with IoT, Algorithms and Research Challenges,» 2021.

[14] M. Di Paolo, «innovationpost.it,» 2019. [Online]. Available: https://www.innovationpost.it/2018/03/08/fog-edge-computing-levoluzione-del-cloud-per-liot/.

[15] V. Lavecchia, «vitolavecchia.altervista.org,» 2018. [Online]. Available: https://vitolavecchia.altervista.org/caratteristiche-e-differenza-tra-cloud-fog-e-edge-computing/.

[16] J. Sheppard, M. Kaufman e T. Wilmering, «IEEE STandards for Prognostics and Health Management,» 2008.

[17] H. Jeong, B. Park, S. Park, H. Min e S. Lee, «Fault Detection and Identification Method Using Observer-Based Residuals,» 2019.

[18] T. Cerquitelli, S. Andolina, A. Marguglio, A. K, P. Petrali e A. Pagani, «Enabling predictive analytics for smart manufacturing through an IIoT platform,» 2020.

[19] I. Christou, N. Kefalakis, A. Zalonis, J. Soldatos e R. Brochler, «End-to-End Industrial IoT Platform for Actionable Predictive Maintenance,» 2020.

[20] S. Panicucci, N. Nikolakis, T. Cerquitelli, F. Ventura, S. Proto, E. Macii e D. Bowden, «A Cloud-to-Edge Approach to Support Predictive Analytics in Robotics Industry,» 2020.

[21] M. Alam e R. R, «Intelligent context-based healthcare metadata aggregator in internet of medical things platform,» 2020.

[22] «en.wikipwdia.org,» [Online]. Available: https://en.wikipedia.org/wiki/Electric_resistance_welding.

[23] «it.wikipedia.org,» [Online]. Available: https://it.wikipedia.org/wiki/Saldatura_a_resistenza.

[24] S. Aslanlar, A. Ogur, U. Ozsarac, E. Ilhan e Z. Demir, «Effect of welding current on mechanical properties of galvanized chromided steel sheets in electrical resistance spot welding,» 2005.

[25] H. Zhang e J. Senkara, Resistance welding: foundamentals and applications, 2012.

[26] Q. Song, W. Zhang e N. Bay, «An Experimental Study Determines The Electrical Contact Resistance in Resistance Welding,» 2005.

[27] P. Sivaraj, M. Seeman, D. Kanagarajan e R. Seetharaman, «Influence of welding parameter on mechanical properties and microstructural features of resistance spot welded dual phase steel sheets joint,» 2020.

[28] S. Aslanlar, A. Ogur, U. Ozsarac e E. Ilhan, «Welding time effect on mechanical properties of automotive sheets in electrical resistance spot welding,» 2008.

[29] C. Vendittozzi e F. Felli, Fondamenti di Metallurgia per l'Ingegneria, 2019.

[30] I. Lum, S. Fukumoto, E. Biro, D. Boomer e Y. Zhou, «Electrode Pitting in Resistance Spot Welding of Aluminum Alloy 5182,» 2004.

[31] X. Zhang, G. Chen e Y. Zhang, «Characteristics of electrode wear in resistance spot welding dual-phase steels,» 2008.

[32] S. Aslanlar, A. Ogur, U. Ozsarac, E. Ilhan e Z. Demir, «Effect of welding current on mechanical properties of galvanized,» 2005.

[33] C. Rajarajan, P. Sivaraj, M. Seeman e V. Balasubramanian, «Influence of electrode force on metallurgical studies and mechanical properties of resistance spot welded dual phase (DP800) steel joints,» 2020.

[34] A. Baskoro, S. Sugeng, A. Sifa, Badruzzaman e T. Endramawan, «Variations the diameter tip of electrode on the resistance spot welding using electrode Cu on worksheet Fe,» 2018.

[35] Y. Li, L. Deng e B. Carlson, «Effects of Electrode Surface Topography on Aluminum Resistance Spot Welding,» 2018.

[36] T. Watmon, C. Wandera e J. Apora, «Characteristics of resistance spot welding using annular recess electrodes,» 2020.