Politecnico di Torino
Department of Electronics and Telecommunications

Institute for Engineering and Architecture

Thesis

For the Degree of

Master of Science in Mechatronic Engineering

Development of a predictive maintenance algorithm for welding gun splash

Domenico Tamborra
(Matr.-Nr.: s260329)

Mentors: Prof. Alessandro Rizzo
Prof. Maurizio Schenone

Supervisor: Ing. Dario Cambiano - ISI-WELDING

Academic Year 2020/2021

Abstract

In parallel with the evolution of the technologies in every industrial field, also the maintenance has been involved in a rapid development and improvement of all that it concerns. Maintenance encloses various and different areas of industrial sectors, from the economy to the security, from the management to the optimization of the use of resources, and therefore over the years it has held a greater and greater importance until it has been established as one of the major strategies to focus on. The increase of reliability and accuracy of algorithms in the Artificial Intelligence field has led several companies to experiment the application of such algorithms in solving maintenance issues.

The ISI-Welding company is a leading company in resistance spot welding. In order to improve the quality of their products and to offer better services to their customers, they had begun a research to implement an intelligent approach to the maintenance of spot resistance welding guns. This thesis work represent the attempt to predict and prevent the occurrence of the splash, a frequent disturbance defined as a spillage of melted metal from the designed welding spot.

A strong basis of the mentioned topics has been gained through the study of the state-of-art of maintenance and welding guns, then a deep knowledge of the system has been acquired analyzing the physical phenomena involved in a welding process. An identification of the model and an estimate of a physical one have been carried out according to the theoretical bases, with good results.

Finally, Machine Learning algorithms have been implemented in order to classify the presence of the splash during the welding process and to predict it. In particular, Neural Net classification and Long Short-Term Memory algorithms have been used.

Results are encouraging and prove the feasibility of this approach, a further collection of larger datasets and the measurement of new critical variables may lead to the development of a complete and exhaustive model of Predictive Maintenance.

Contents

1 Introduction **8**
1.1 ISI-Welding . . . 8
1.2 Objective . . . 9

2 State of art **10**
2.1 State of art of welding guns . . . 10
2.1.1 Electrical part . . . 12
2.1.2 Pneumatic part . . . 14
2.1.3 Mechanical part . . . 15
2.1.4 Handling system . . . 18
2.1.5 Cooling system . . . 19
2.2 State of art of maintenance . . . 20
2.2.1 Lean manufacturing . . . 20
2.2.2 Maintenance policies . . . 22
2.2.3 Maintenance strategies . . . 22
2.3 State of art of machine learning . . . 26
2.3.1 Classification Problem . . . 27
2.3.2 Regression problem . . . 29

3 Model **30**
3.1 Data structure . . . 30
3.2 ARX, ARMAX and OE . . . 31
3.2.1 Polynomial model . . . 32
3.2.2 ARX model structure . . . 32
3.2.3 ARMAX model structure . . . 33
3.2.4 OE model structure . . . 34
3.2.5 Data processing . . . 35
3.2.6 Results . . . 37
3.3 NARX Model . . . 42
3.3.1 Results . . . 43
3.4 Physical model . . . 49

4 Splash definition **54**
4.1 Splash classification . . . 56
4.1.1 Statistical descriptors . . . 58

5 Splash prediction **63**
5.1 Splash prediction with classification . . . 63
5.2 Splash prediction with time-series algorithms . . . 65
5.2.1 LSTM algorithm . . . 65
5.2.2 Forecasting with LSTM algorithm . . . 66
5.2.3 Forecasting with NARX algorithm . . . 72

6 Conclusion **75**

6.1 Future developments 76

List of Figures

1 Decomposed fulcrum spot welding gun 11
2 Flaps and shunts 12
3 Arms, electrodes and cooling system 13
4 Transformer 13
5 Signal stip detail 14
6 Pneumatic balancing group 14
7 Equalizing system 15
8 Fixed support 16
9 Fixed articulation 16
10 Mobile articulation 17
11 Consolle 18
12 Handling system 19
13 6 S flow 21
14 Machine learning flow 27
15 Neural network simple graphical structure 28
16 Example of the system to identify 32
17 Example of an ARX model block diagram 33
18 Example of an ARMAX model block diagram 34
19 Example of an OE model block diagram 35
20 Example input in a short circuit welding point 36
21 Example output in a short circuit welding point 36
22 Auto-correlation and cross-correlation analysis 37
23 RMSE analysis of ARX model 39
24 Measured and estimated output comparison 40
25 Measured and estimated output (single input models) 41
26 Measured and estimated output (double input models) 41
27 NARX architecture 42
28 Comparison between the real and open-loop estimated output 43
29 Comparison between the real and closed-loop estimated output 44
30 Comparison between the real and closed-loop estimated output without blanking 45
31 Comparison between the real and open-loop estimated output 46
32 Comparison between the real and open-loop estimated output, without splash 47
33 Comparison between the real and open-loop estimated output, with splash 47
34 Comparison between the real and closed-loop estimated output, with splash 48
35 Weld core of metal sheets during a welding process 49
36 Comparison between estimated and real model (short circuit case) 51
37 Comparison between estimated and real model (real welding case) 52
38 Comparison between physical model and real outputs 53
39 Spillage of melted metal 54
40 Optimal welding point 55
41 Welding point affected by splash 55

42 Trend of the neural net training 57
43 Training and test confusion matrix 57
44 Confusion plot of splash classification 58
45 Skewness and kurtosis graphically explained 60
46 Confusion plot of splash classification with statistical descriptors 61
47 No-splash points, according to WMS 62
48 Confusion plot of splash prediction with classification 63
49 Confusion plot of splash prediction with classification 64
50 Recurrent neural network 65
51 Long short-term memory network 66
52 Voltage prediction through LSTM algorithm, point without splash 66
53 Voltage prediction through LSTM algorithm, point with splash 67
54 Voltage prediction through LSTM algorithm without blanking 68
55 Voltage observations and 4 ms prediction, without blanking 69
56 Real and forecasted state of the voltage 70
57 Voltage observations and 8 ms prediction, without blanking 71
58 Real and forecasted state of the voltage 71
59 Voltage prediction through NARX algorithm, point with splash 72
60 Simulation of a real welding process, 5 ms prediction 73
61 Simulation of a real welding process, 8 ms prediction 74

List of Tables

1 Failure classification 23
2 Benefits of the predictive condition-based maintenance 24
3 ARX, ARMAX and OE orders and delays 37
4 ARX residues analysis results 38
5 ARX, ARMAX and OE orders and delays for both input conditions 40

1 Introduction

In these years characterized by a rapid development of technologies, companies have been challenged to radically change their habits and to find modern solutions that fit the context. Being constantly updated is essential to keep competitiveness high and it is a fundamental quality for customers' eyes.

Hand in hand with the development of data science, predictive maintenance is one of most interesting topics of recent times. The main promise is to allow convenient scheduling of corrective maintenance and to prevent unexpected equipment failures, saving goods, money and production time.

Moreover, in this particular and historical period of global pandemic, companies are unwilling to buy new goods rather than maximize the useful life of their own industrial machines.

The estimation of the state of health of the machines is now an Artificial Intelligence issue, as it can lead to several benefits and profits such as cost saving, safety, resources availability. The management of larger databases, the improvement of the accuracy of sensors and measurements, the high computational power and the high CPU speed set up the perfect environment for the using of Machine Learning algorithms.

More and more companies are trying to break down the barriers between sellers and customers. An innovative maintenance methodology can improve the relationship with the customer going beyond the sale of the product: a predictive maintenance system would allow the selling company to closely deal with the working life of their machine and to act promptly in case of need, eliminating external maintenance interventions and increasing the customer loyalty.

1.1 ISI-Welding

ISI-GF EQUIPMENT (WUHAN) CORP., LTD. is a professional manufacturer of intelligent robot welding equipment and supplier of integrated welding robot technology, with major business of research, development, production and distribution of intelligent robot welding products and welding robots. Since its establishment in 2007, the company has been focusing on the application of robot welding technology to automobile manufacturing industry, making it one of the few manufacturers in automobile industry in possession of integrated solution for intelligent welding robot for customers. As a manufacturer of intelligent robot welding equipment and assembler of welding robots, they are committed to become a first-class "supplier and service provider of all-in-one solution for intelligent robot-based welding operation", that is, design professional intelligent solution adapting to customer's special requirements on welding assembly, supply corresponding intelligent robot welding equipment, complete welding procedures and provide the customer with satisfactory after-services. By business optimization and integration, the company is transforming to a high-tech enterprise able to deliver the customers with all-round services from technical solution, design, processing, production, manufacture, installation, delivery to training and consultation. ISI have constructed industrialized production base in Wuhan, where it's equipped with complete production, monitoring, testing and experimental installations for production and manufacture of welding control system, integrated welding machine, MF welding system, robotic automatic welding system, nonstandard welding fixture and its packages and complete automatic welding production line. ISI-GF EQUIPMENT was delisted in August 2015 and acquired ISI-Italia (Original Italian GF Welding S.p.A.), a company specialized in welding technology with a history of 50 years, acquiring world-leading robot welding technology and development capacity, and effectively expand to global market. Acquired by SI-GF Equipment (Wuhan) Co., Ltd, ISI-Italia (Former Italian GF) is a professional firm of welding technology, a producer among few world players able to supply complete welding technology, with all know-how derived from experiences for over 50 years, full line of core products are independently developed and designed, also, it's an exclusive supplier of FIAT Italia, key supplier of Volkswagen Deutschland, key supplier of Renault France, as well as a supplier for automobile engine manufacturers, including French PSA. Wuhan ISI-GF Eagle Automotive Equipment Co., Ltd, a controlling subsidiary of SI-GF Equipment (Wuhan) Co., Ltd, was established in 2015, the business scope of which includes: development, design and manufacture of automatic production line as well as mechanical electrical equipment; installation and refitting of mechanical equipment; manufacture, wholesale and retaining of position apparatuses, fixtures, molds and gages (Special equipment not included); development and technical transfer of automotive assembly technology; wholesale and retaining of automotive assembling tools, wires & cables, electronic products, metallic materials, steel structure members, automotive parts, integrated mechanical & electrical products and accessories.[1]

1.2 Objective

The objective of the ISI-Welding proposed internship is to develop a model for the predictive maintenance of their welding guns.
The work is focused on two important welding guns characteristics:

- **Splash**, a disturbance defined as the spillage of melted metal out of the working area;

- **Electrode dressing**, defined as an electrode reshaping by milling in order to remove the pollution.

From the collaboration of Politecnico di Torino with the ISI-Welding Company, a team of two graduate students supervised by the R&D director Eng. Dario Cambiano has been created. This team has shared the first part of the work, made up by the study of the state-of-art of welding guns, maintenance and machine learning and by the research of a suitable model of the system. Once a solid knowledge has been acquired, each student performed his own work on the two mentioned topics. In particular, this thesis work is focused on the splash disturbance.

2 State of art

2.1 State of art of welding guns

The spot resistance welding is a welding methodology 'luckily' born in the XIX century. During a physics lesson at the Franklin Institute in 1877, the teacher Elia Thompson was illustrating to his student a simple electrical circuit, an induction coil with some capacitors on the secondary winding. Once the capacitors were charged, he tried to short circuit the primary winding, the current melted the connected ends of the wire, joining them together. This was the first resistance welding of history. Since then, resistance welding has made enormous progress up to current technology.

The welding gun is an equipment provided by mechanical and electrical elements. In order to obtain a resistance welding spot, this machine should be able to compress the sheets to be welded and to release high currents that melts the metal through Joule effect. An actuator generates the compression force, a transformer generates high currents (tens of kA) with low voltages (15-20 V) from mains voltage (380 or 500 V). In order to avoid the melting of the gun electrical components and to better cool the weld core, the welding gun is provided by a cooling system (usually the refrigerant fluid is water). The welding guns can be manual or robot. They obviously have the same mechanical and electrical characteristics, but the robot welding gun is the control unit of a robot and it automatically moves and performs a sequence of welding spots. The manual welding gun has to be carried to the working positions by a human operator. Welding guns can also be classified as:

- fulcrum welding gun
- slider welding gun

The difference consists in the arm movements toward the welding spot: a welding gun is provided with a fixed arm and a mobile one. In the fulcrum gun both arms are hinged to the same fulcrum, the mobile arm rotates around this fulcrum to reach the welding spot. In the slider gun the fixed arm is connected to the frame, the mobile one performs a translation movement in order to compress the sheets in the designed spot.

A fulcrum welding gun is composed from the following functional groups :

- Electrical part:
  - **Shunt (01)** form the electrical connection between transformer and welding part.
  - **Welding part (02)** includes electrode (tips), electrode holders and arms.
  - **Transformer (03)** inserted between the brackets and equipped with all the sensors to measure current and voltage.
- Pneumatic part:
  - **Equalizing system (04)** with valve group.

  - **Actuator cylinder** just for welding guns with pneumatic implementation.

- Mechanical part:
  - **Fixed support (05)**.
  - **Fixed joint (06)**.
  - **Moving joint (07)**.
  - **Robotic arm connection (08)**.

- Handling system:
  - **Electric motor (09)**.

- Cooling system:
  - **Hose fittings**

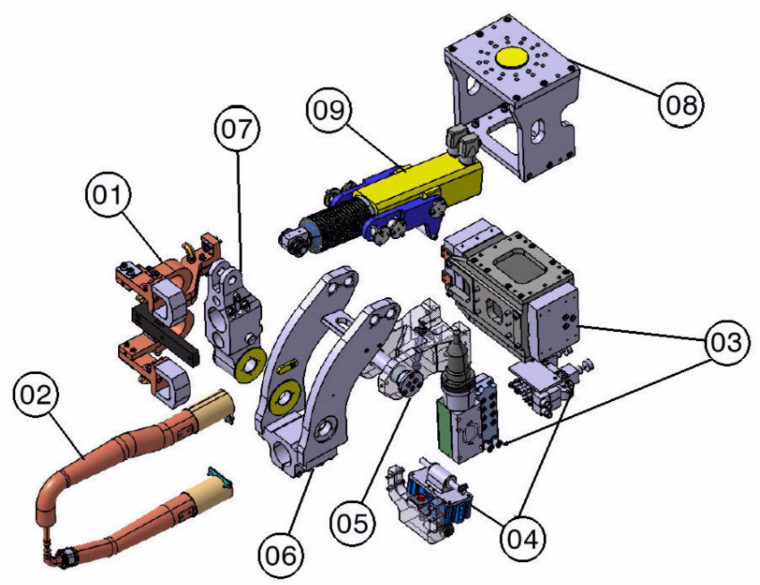


Figure 1: Decomposed fulcrum spot welding gun

2.1.1 Electrical part

Shunt
Flaps group is composed from rigid connections realized with electrolytic copper (a) by fusion or using commercial slabs, they are screwed on the transformer case and are called fixed shunts. The shunts are individually electrically isolated using a bakelized canvas (b) and also all the bushings and washers are insulators. Then there are flexible connections, the lamellar packs (c), they are called shunts and are silver-plated (to improve conductivity) copper bundles. Flaps guarantee electrical continuity allowing arms movement. Flaps are connected to the arms through clamps called brides (d), these components realize a mechanical, electrical and fluid junction at the same time. For this reason, surface tolerances (e) for these components are really restrictive. On the same surfaces, the hole for the cooling system is placed.

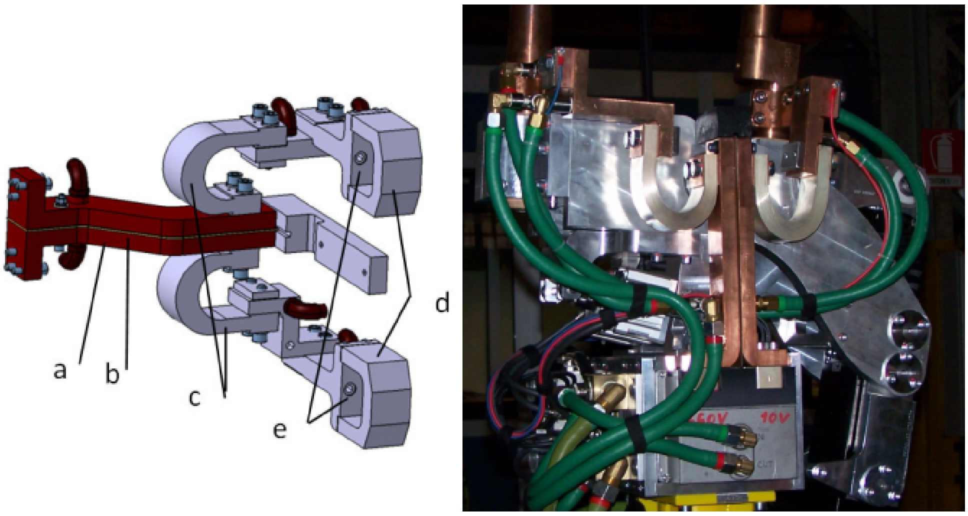


Figure 2: Flaps and shunts

Welding part
Welding part is composed by the arms (a) realized in copper alloy (CuCrZr) starting from commercial cylindrical sections. These components have both structural and electrical function because they have to allow force transmission to the electrodes (b) and guarantee electrical continuity between shunts and metal sheets. On the arm is placed the electrode holder (c) according with standards required by car manufacturer companies. On the electrode holder is positioned the electrode, often with a conical section acts to maintain the cooling fluid and to easily perform replacement operations. Inside the arm and the electrode there is the cooling circuit realized in copper (d). Fluid is managed by a brass pawl (e) with o-rings (f), the cool water travel inside the arm while the heated one has an external path. The caps (g) close the hole used for liquid insertion.

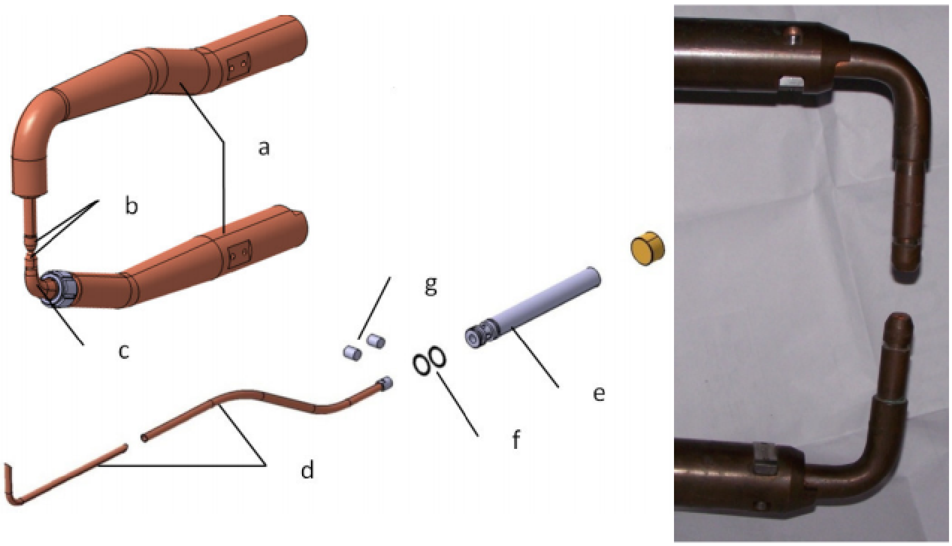


Figure 3: Arms, electrodes and cooling system

Transformer
The transformer (a) is correlated with 4 brackets (b) that form its cage. The transformer cage is both a support element and a robot coupling. Flaps are connected on a side of the transformer (c), on the other side there is the signal strip and power connection.

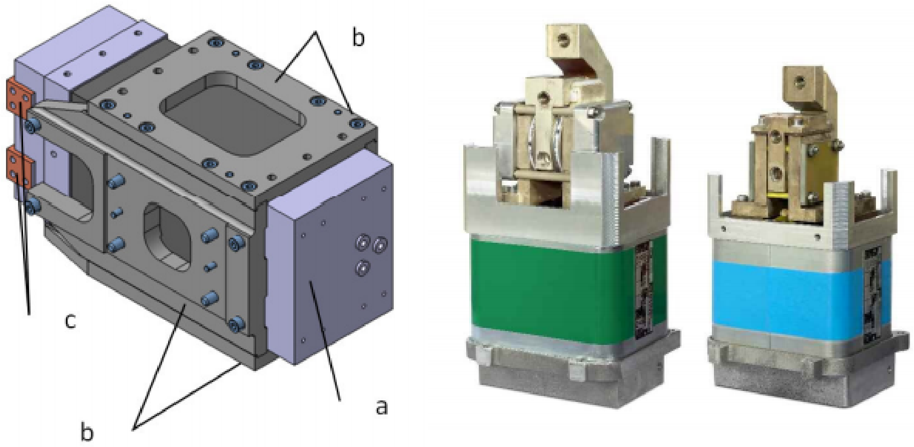


Figure 4: Transformer

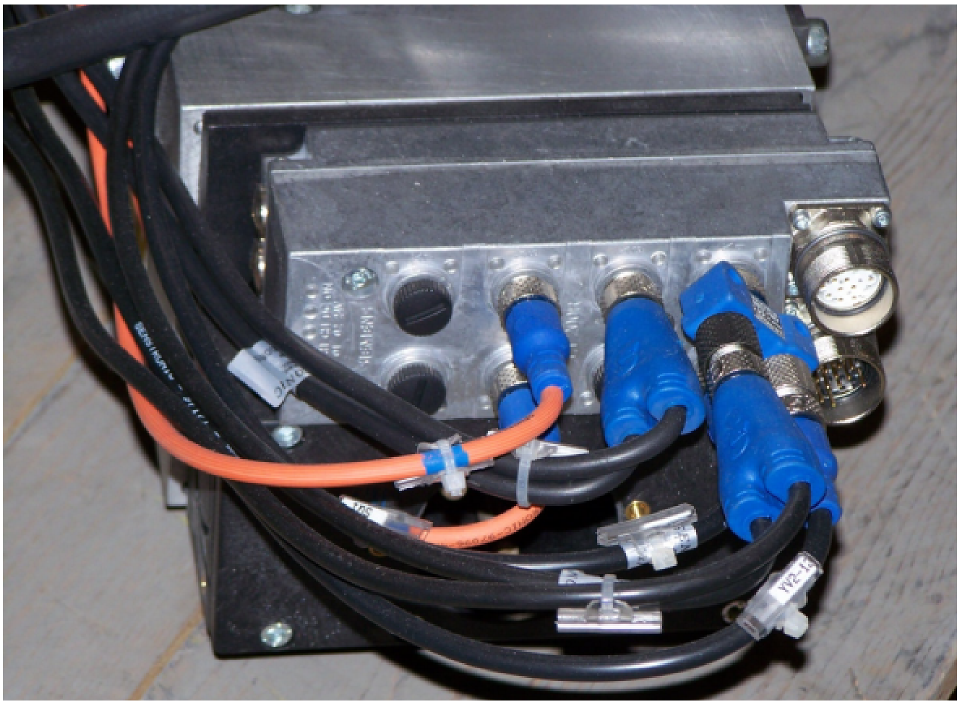


Figure 5: Signal stip detail

2.1.2 Pneumatic part

Pneumatic equalizing group
The pneumatic balancing group consists of one or two pneumatic cylinders (a).
The cylinders are hinged on one side to a bracket (b) connected in turn with the lower support, on the other side the cylinders are connected to the fixed support. The balancing assembly also carries an end pad stroke (c).
Balancing requires a control valve group that also contains manometer indicators. This valve assembly is typically placed on the transformer bracket assembly in the space-saving position.

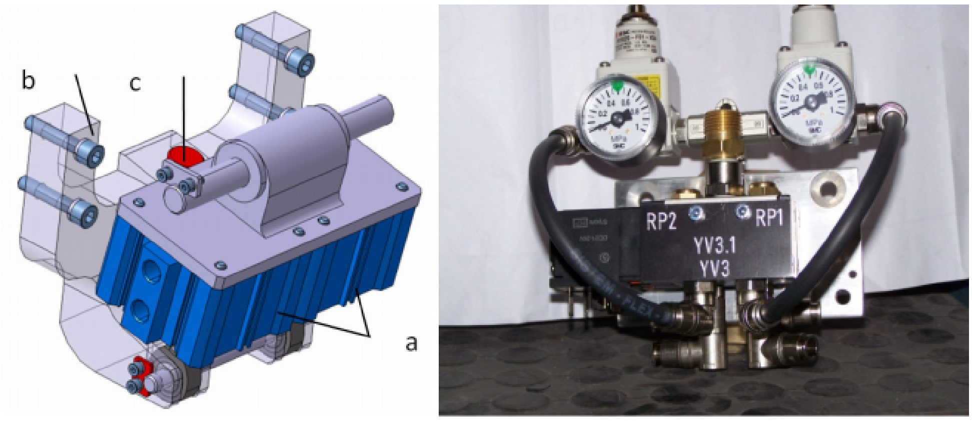


Figure 6: Pneumatic balancing group

In recent times, in order to lighten the welding gun, the pneumatic group has been removed and replaced with a software function able to equalize the force on each tip. An equalizing system compensates for welding conditions in which the closing weld tips are offset from the plane of the workpiece. As one of the tips first touch the workpiece, a force is created that slides or rotates the gun to a position that centers the gun tips about the workpiece. [2]

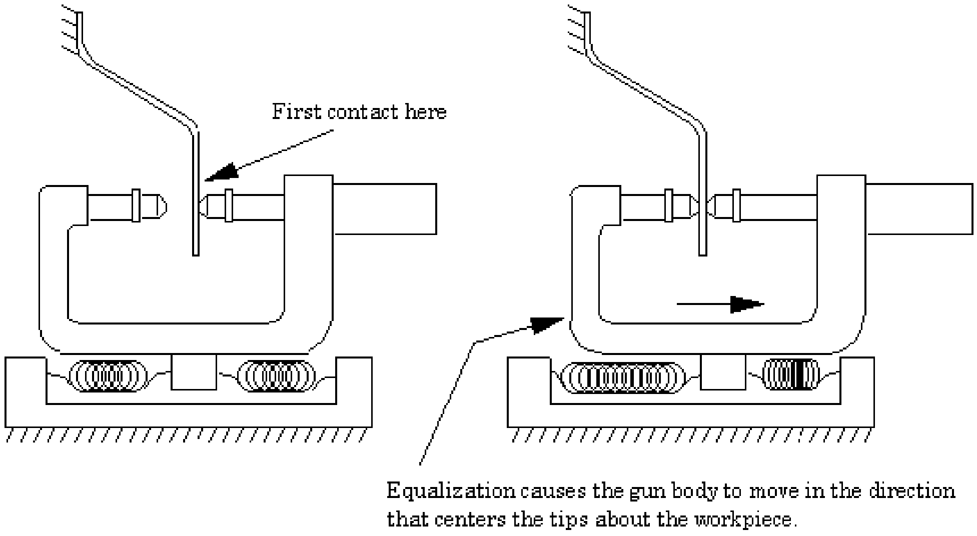

First contact here
Equalization causes the gun body to move in the direction that centers the tips about the workpiece.

Figure 7: Equalizing system

2.1.3 Mechanical part

Fixed support
It is the main structural element of the fulcrum clamp on which all the forces and the moments are discharged.
The fixed support is made of aluminum alloy, obtained by fusion, the support is equipped with the pivot (a) on which the joints are hinged. The support is connected at the rear to the transformer cage. It also contains the screw (b) that allows the adjustment of the balance and the seat for the cylinder connection pivot balance (c).

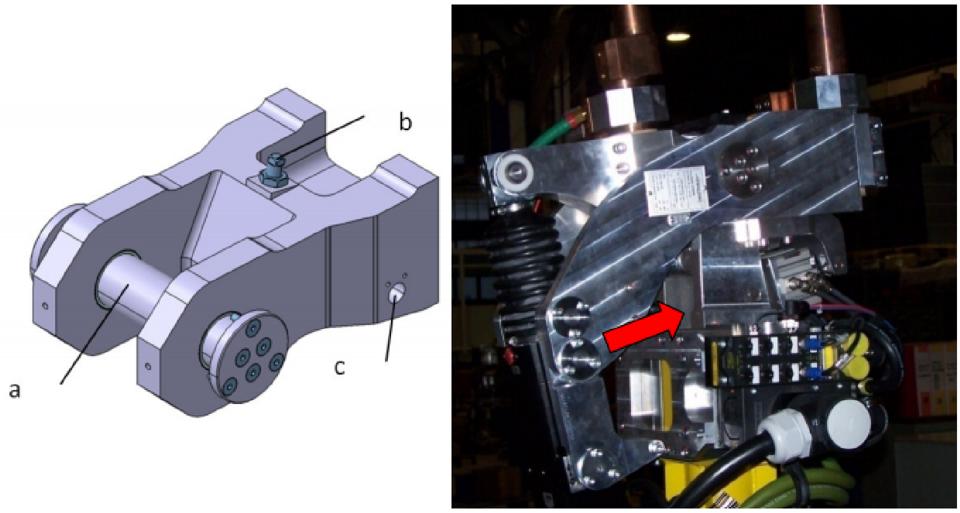


Figure 8: Fixed support

Fixed articulation
It is an assembly of components in Ergal 7075 consisting essentially in two sides (a) that carry the seats of the fulcrum (b) and the seats of the connections to the handling system (c). In the lower part a support (d) is connected to the sides. In the upper part there is a reinforcement (f) which prevents the sides from twisting. The plate (i) is the guide for moving the mobile arm.
To avoid the rotation of the arm during the application of the welding force it has been added an anti-rotation system consisting of a leveling on the arm and a corresponding plate screwed on the back. The figure on the right highlights the anti-rotation system (e), the insulating bush (g) and the adapter bush (h) for the different arm diameters.

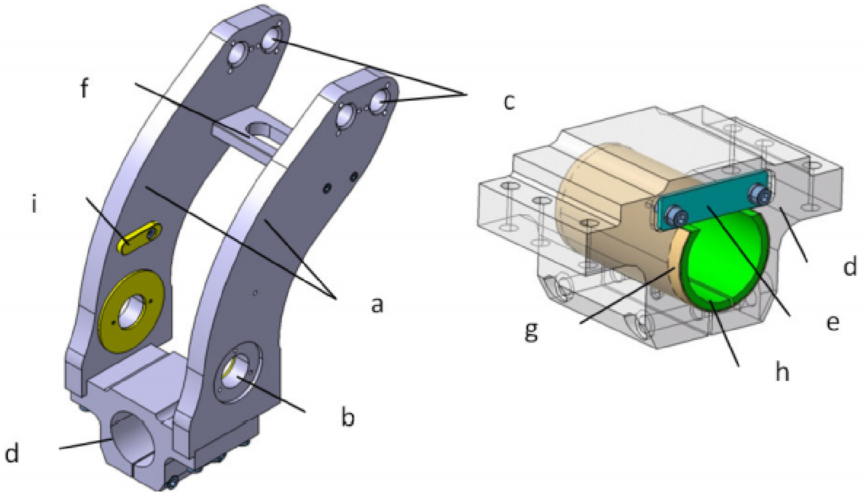


Figure 9: Fixed articulation

Mobile articulation
The mobile joint is inserted at the bottom on the main fulcrum while at the top it is

inserted in the pivot of the tenon by means of a ball joint that allows small misalignments of the pivot axis with respect to the motor or cylinder rod.

The figure shows in yellow the insulating washers (a) in Ertalyte TX, fixed with pins for prevent its rotation; these serve as insulators and help eliminate gaps between the joints.

To uniquely guide the arm, pads of polymer (b) are added on the sides of the joint, they slide on the guides on the fixed joint.

The closing of the arm (c) in the clamp is carried out in the same way as that of the lower arm.

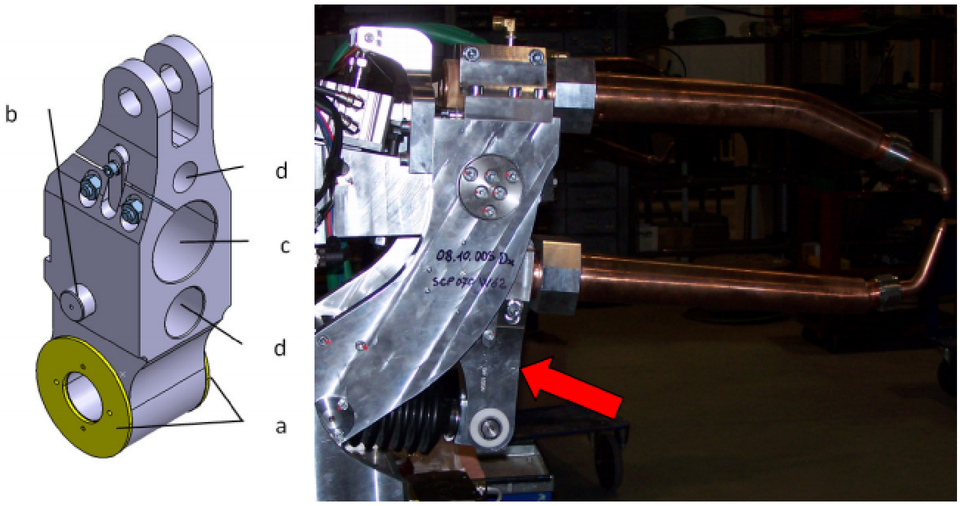


Figure 10: Mobile articulation

Consolle
The consolle or robot attachment is a structure in Ergal 7075 that is used to connect the robot wrist to the transformer cage. The shape of the robot side flange depends on the manufacturer's standards robot (in the figure a Comau standard for Smart series robots), the shape of the other plate must instead mate with the brackets of the transformer cage. The most common variants are two. The first variant (I) is relating to an upper robot attachment and it is made with screwed plates. In that case it is necessary to override the actuator with the console. The second variant (II) with welded plates is used when the attachment is lower or rear. Attachments can present different distances and angles between the plates. The yellow disk (a), visible in the figure, is a centering device. Sometimes a lateral attachment is also possible.

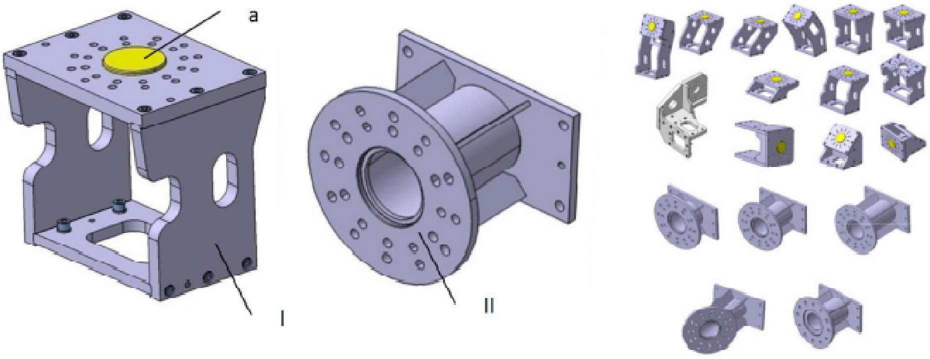


Figure 11: Consolle

2.1.4 Handling system

The most used handling systems are electric or pneumatic. The electric actuator stem (a) is connected to the tenon (a component not shown which carries the seat of the attachment pin to the mobile joint), the engine is connected to the fixed joint through a subframe (b) while the pneumatic cylinder to the other components in the same way but it does not require auxiliary frames. However, the cylinder obviously has the so-called 'bar kit' (c) that is the system of fittings, valves and filters for the management of compressed air. The pneumatic cylinder also has micro brackets (d) for stroke adjustment if it does not have a servo control.

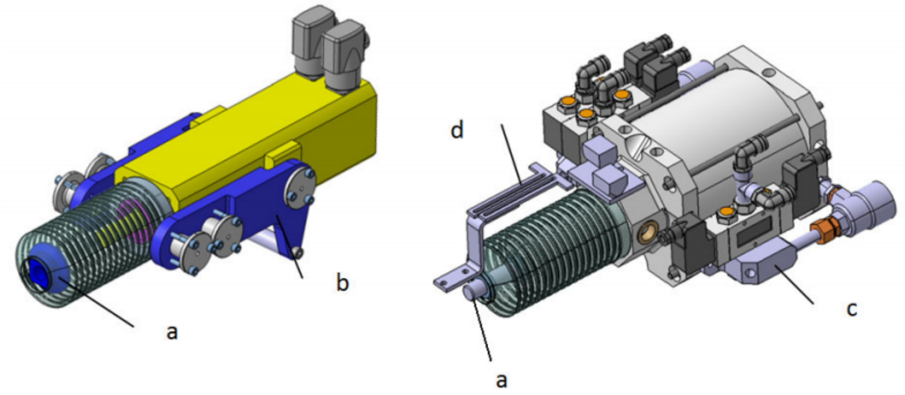


Figure 12: Handling system

2.1.5 Cooling system

The cooling system consists of two main circuits. In the clamps with transformer a 50 Hz a circuit cools one arm and the flaps, the other one arm and the transformer. In the clamps MF the cooling circuit of the transformer is independent, this is because the transformer MF requires more heat removal.[3]

2.2 State of art of maintenance

Maintenance is a set of technical, operational and managerial actions with the aim of guarantee the availability, cost-effectiveness and safety of systems and the optimal use of resources. A first phase of the organizational development of maintenance can be located in the ’60s-’70s, when the importance of a maintenance planning and improvement was felt in sectors with growing market, such as steel, chemical, petrochemical and aeronautical one. In particular, thanks to the aeronautical industries, the reliability theory was developed. This probabilistic theory was based on mathematical and physical theories and it was aimed at estimating the remaining life of a component. A second important phase, during the ’80s, was characterized by the overcoming of the maintenance endorsement, for example transferring maintenance resources to production departments and teaching the basics of maintenance to human operators. This path led to a third phase in which production appropriates the maintenance culture until the complete integration of production and maintenance strategy.

2.2.1 Lean manufacturing

Lean manufacturing, or lean production, is a production method derived from the Toyota strategy of the 1930 and was defined from Womack and Jones as ’the way to do more and more with less and less’, the way to give to the customer exactly what he wants using less effort, less time, less equipment and less space. This way of thinking can be resumed in 5 key principles:

- **Value:** specify the value of the product as it is desired from the customer.
- **Value Stream:** identify the value stream for each product.
- **Flow:** make the product flow continue, without interruption.
- **Pull:** introduce pull between steps to make the flow continue.
- **Perfection:** improvement

The continuous improvement is then realized with the kaizen philosophy (composed of the 6 S principles).

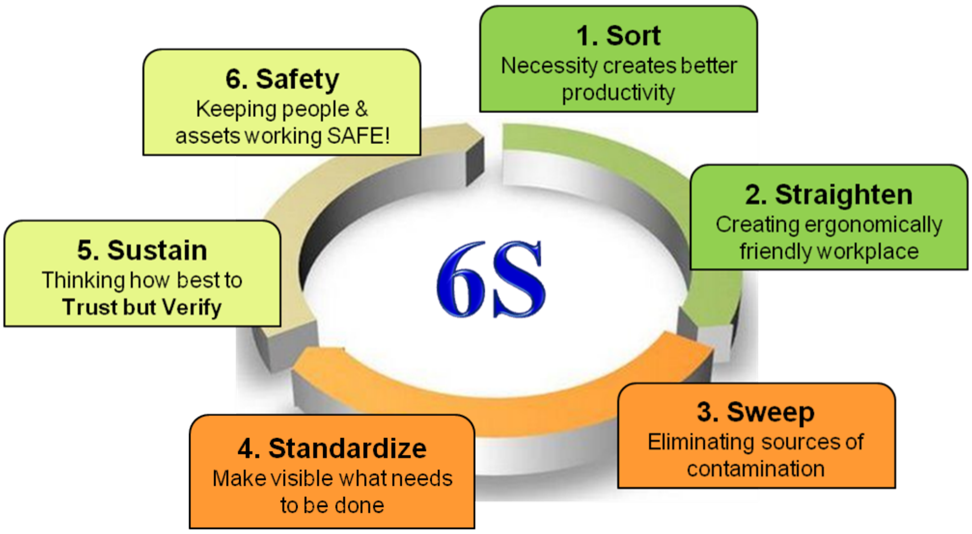

1. Sort
Necessity creates better productivity
2. Straighten
Creating ergonomically friendly workplace
3. Sweep
Eliminating sources of contamination
4. Standardize
Make visible what needs to be done
5. Sustain
Thinking how best to
Trust but Verify
6. Safety
Keeping people & assets working SAFE!
6S

Figure 13: 6 S flow

The modern concept of maintenance arises from 2 fundamental assumptions:

- **Increasing automation:** human operator now-day is the manager and supervisor of the machine.
- **Growing competitiveness:** the imperative is to serve excellence to the customer.

In this contest, the entire production process is divided in elementary productive unit: the mini-factory. Every mini-factory carries out only one of the transformations that lead the raw materials to the final product. The process is managed independently: the mini-factory has full responsibility also in the search for suppliers and any other customers. Human operator in this context is the supervisor of an activity and he has to be able to guarantee the quality through self-certifications. The self-certification defines the critical variables of the process: the aspects to which particular attention must be paid to have a satisfactory quality.

This category also includes the parameters that most influence customer satisfaction even if they are not considered, from the producer, fundamental for process quality.

The fundamental concept in this philosophy is "The Empowerment": exploit as much as possible the resource of the human intellect. The Empowerment provides to transfer more and more elementary functions to the driver of the machine. For this reason, modern maintenance starts with cleaning the machine. In this way the worker can learn how the machine is done, how it works and where are placed the critical aspects. Over time, by cleaning, the human operator can learn about the places that get dirty the most but also can notice, with the experience, of parts that are deteriorating and wearing out. The worker become an integral part of the maintenance system, he will suggest to expert the critical issues to be analyzed.

After this step, obviously, there are: the establishment of an information system with suitable diagnostic tools, the planning of cyclical interventions and the optimization of the life cycle cost.

The goal of the modern maintenance philosophy is the preservation of the heritage while avoiding temporary actions that damage machinery in the long term. Also increased availability and quality are values to be pursued, new technology must be viewed with distrust especially if it has not been tested for a long enough period. History is full of example of how new technologies, too reckless, have brought disastrous and sometimes even catastrophic results.

The last but not least point is costs reduction which is leading to an increasing 'Outsourcing'.[4]

2.2.2 Maintenance policies

The maintenance policy indicates the overall attitude that the organization assumes in relation to maintenance problems, which can then be explicit in the use (depending on the departments, the single machine, the economic convenience etc.) of various strategies. One of milestones of maintenance approach was the *Total Productive Maintenance*, developed in Japan in fairly recent times, defined as "production maintenance carried out by all workers of the company organized in small groups of activities"[5]. It is a comprehensive approach to organizational issues with a view to improving the performance of production equipment and plants, which takes into account the Japanese matrix and the application experiences made in the Italian industry. The main innovation was bringing under the responsibility the coordinator of a production segment at the operational level also the maintenance line and quality control. So TPM's main contribution to maintenance theory is given by the attempt to break down the existing demarcation line, within a company, between maintenance and production departments. In this context, the TPM acknowledges the existence of several maintenance situations which may require different techniques to achieve a good result, and consequently it uses different methodologies which can differ from plant to plant or from machine to machine, provided that they are effective in a given situation. Many of the strategies used are certainly not new: what is innovative is the Japanese culture, the commitment it provides for all employees. Business maintenance policy optimization should be pursued in the context of improving business profitability and the service provided and, in particular, the continuous improvement in operating income. This improvement is the expression of a close synergy between maintenance and production which takes the form of production maintenance.

2.2.3 Maintenance strategies

The maintenance activity aims to obtain a certain continuity of the production process, in the past this objective was pursued through operational and functional redundancies or applying an aggressive program of review and replacement of critical systems. All these approaches have proved to be partially inefficient, as redundant systems and excess capacity freeze capital that could be more profitably used for productive activity, while a political revision excessively prudent has proved to be a rather expensive method to obtain the required standards. Maintenance has therefore transformed from operational repair activities to a complex management system with the point of preventing failure.

Before discussing the existing maintenance strategies, it is important to focus on the different kinds of failure that a system can deal with.

Table 1: Failure classification

Failure type	Description
Catastrophic failure	A condition of sudden and complete cessation of operations and a total deterioration of functions.
Sudden failure	A condition of accelerated degradation of both material and performance, which results in a partial weakening of functions.
Imminent failure	A condition of perceptible degradation of the material in the presence of a serious deterioration in performance.
Incipient failure	A condition in which the use of appropriate means of investigation allows to identify the first signs of degradation of the material, without the user experiencing any change in the performance of the machine.
Conditional failure	A condition of pre-alert in which it has not yet occurred a deterioration neither of the material nor of the performance but such that, if the situation persists, it will inevitably lead to a functional failure.

Each strategy is composed to a set of actions suitable for a specific failure. Basically, the failure on which it is decided or it is possible to act determines the strategy to follow.

Breakdown maintenance
Breakdown Maintenance is certainly the most spontaneous and simple way to work: maintenance action is taken when the failure occurs. In the presence of non-critical and easy-to-replace systems at low cost, it is convenient wait for the failure to occur rather then applies a more expensive preventive approach. Unfortunately, this strategy has many questionable aspects: a serious and unexpected failure on a component may have deleterious consequences on other elements of the system, compromising their functionality with an additional amount of costs, moreover unscheduled repairs often take a long time to obtain spare parts and assign the appropriate technician, stopping the production and poorly employing human operators. Finally, a sudden or catastrophic failure is a condition that a good maintenance activity should avoid a-priori.

Preventive maintenance
Preventive Maintenance is based on the belief that the average life of some component is determinable and that it is possible to anticipate the failure of a system (machine or production line), pre-defining the moment of intervention, usually replacement, depending on the expected lifetime of the component itself. This concept was a great success in the 1960s and 1970s with the spread of the reliability theory, because it gave a basis of scientific nature to the maintainers. It is a type of maintenance that is one step higher than the previous one, because in this case the mechanical system is still working but its performance deteriorates to the state of imminent failure. There are two philosophies to implement a failure avoidance:

- Condition-based, that promotes maintenance only when necessary by means of a shallow observation of the system and the detection of the deterioration,

- Time-based, that schedule the interventions at constant interval on the basis of reliability, safety and performance.

The weak point of this strategy is that the reliability theory is a probabilistic theory, so a failure can happen also before the scheduled part replacement, but mainly there are no chances to increase the mean time between two subsequent failures of the system.

Predictive maintenance
A modern view of maintenance problems led to the use of non-destructive techniques for testing systems for the purpose of identifying with a consistent advance the presence of faults, so it is possible to schedule a review only when the condition of the machine determines its necessity. This maintenance strategy does not use probabilistic methods for making a prognosis of the failures, but it uses the trend of tracked parameters to predict potential failures. This is the predictive maintenance: a diagnostic process that, by providing information on the health status of machine allows to plan revisions based on the actual conditions of the components rather than on the operating time. Unlike the earlier described condition based preventive maintenance, this new philosophy has significant implications on design: in fact, to reduce to minimum passive times due to frequent checks, the mechanical system should be equipped with a whole series of devices necessary for the determination of the status efficiency of components.

Table 2: Benefits of the predictive condition-based maintenance

Safety	Predictive maintenance allows machine downtime before reaching critical condition
Increase in availability, lower costs maintenance	The intervals between two successive revisions may be increased. Downtime can be reduced by preparing maintenance resources
Better chance of negotiation with manufacturers	Because the conditions are measured on new machines, at the end of the warranty and after the review it is possible have some comparison data
Better relationships with customers	Knowing in advance when a failure will occur, it is possible better organize production
Opportunities to design better future plants	The experience properly collected in historical files can be useful for this purpose

The limit of predictive maintenance can be identified as being failure-oriented: it is more effective than traditional approaches but leaves wide areas of improvement in terms of reliability and cost reduction. This strategy tries to provide the operator an sufficient warning alert to organize the necessary repairs and the downtime. This depends, of course, on the monitoring program and the time needed to obtain the results of the analyses: if this time is large, an incipient failure conditions may transform in imminent failure one, bringing the system into much more worrying conditions.

Proactive maintenance
All these maintenance strategies can be defined as 'reactive' strategies. In the Proactive (or productive) Maintenance, the term 'proactive' is opposed to the concept of reaction, in the sense that it refers to an action that takes place before the critical event. It is a pre-alert activity that is carried out before any damage relating to the equipment or performance of the system, a series of actions with the aim of correcting those conditions which may lead to deterioration of the system. Instead of analyzing material or performance alteration to evaluate incipient or imminent failure conditions, the proactive maintenance is proposed to detect and correct abnormal values of primary causes of failure that could lead to conditions of operational instability, the so-called 'failure roots', and report that first level of malfunction, the 'conditional failure'. This maintenance practice is the first line of defense against the degradation of material (incipient failure) and the consequent weakening of the performances (imminent failure) which finally lead to the breakdown. Moreover, an intervention with such advance make possible to avoid the occurrence of secondary failures that may arise on the elements adjacent to that in examination (for example because of vibrations). Summing up, the proactive maintenance requires the following actions:

- monitoring of the key parameters indicative of the health of the system (failure roots),
- definition of threshold values, that is the maximum acceptable values for each parameter,
- recognition and interpretation of any outlier of these key parameters, which indicate some instability in operating conditions,
- Specification of the methods to be used to correct primary failure causes and restore system stability.

The evaluation of the failure roots is not always possible, sometimes there are no ways to detect them or sometimes it is too expensive. Implementation of the maintenance policy requires design criteria based on the logic of minimizing the overall cost. The first step is analyzing a specific failure mode and verifying the existence of measurable signals that can help its detection. If the signal exists, it is possible to perform a predictive condition-based maintenance by monitoring the degradation of the component. If the signal does not exist, then the analysis moves on the theory of reliability and the estimated life of the component. If there are enough information about these topics, a preventive maintenance can be implemented activating planned inspections or performing replacements at scheduled times. When there is no signal e no estimated life of the component, the breakdown maintenance is the only possible strategy.[6]

2.3 State of art of machine learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.
Machine learning starts with the collection of a great amount of data, they can be measured on a real plant or simulated using an identified model of the system.
In this way the algorithm can look for common patterns in data and make better prevision in future based on the examples that we provide. Machine Learning algorithms can be categorized as supervised or unsupervised.

- **Supervised machine learning algorithms:** the dataset provided to train the algorithm are labeled examples. The learning algorithm produces an inferred function to make prediction about the output values and compare its output with the given label to accordingly modify the parameters inside it.

- **Unsupervised machine learning algorithms:** they are used when the information is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system does not figure out the right output but can find out hidden patterns from unlabeled data.

- **Semi-supervised machine learning algorithms:** between supervised and unsupervised learning, since this kind of algorithm uses both labeled and unlabeled data for training, typically a small amount of labeled data and a large amount of unlabeled. This method is able to improve learning accuracy. Semi-supervised learning algorithms are chosen when labeled data requires a great number of resources to be collected, while acquiring unlabeled data generally does not require additional resources.

- **Reinforcement machine learning algorithms:** this is a learning methodology able to interacts with its environment by producing actions and discover errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. A simple reward feedback is required for the agent to learn which action is the best, this is known as the reinforcement signal.

Machine learning allow the analysis of massive quantities of data and, to be trained properly, it may require additional time and resources.

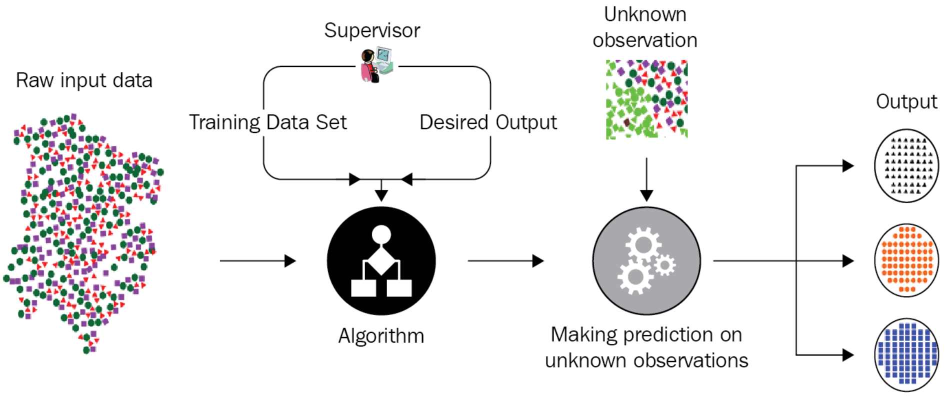


Figure 14: Machine learning flow

2.3.1 Classification Problem

In statistics, a classification problem consists in the identification of a set of categories. This approach is widely used to image recognition applied on autonomous vehicles, for example. The success of this kind of algorithm is entrusted by the presence of a sufficiently wide data set, most of the time labeled, that allows the solution of a supervised problem. In the terminology of machine learning the corresponding unsupervised procedure is known as clustering and involves grouping data into categories based on some parameters that highlights similarity or distance functions. Often, the individual observations are re-arranged into a set of quantifiable properties known as explanatory variables or features. These properties may be categorical, ordinal, integer-value or real-valued.

Other classifiers work by comparing recent observations to previous observations by means of similarity or distance function. An algorithm that solves a classification problem is a classifier. The term 'classifier' sometimes also refers to the mathematical function implemented by a classification algorithm to map the data into a category. Classification and clustering are examples of the more general problem of pattern recognition, which is the assignment of some sort of output value to a given input.

A subclass of classification is probabilistic classification. This kind of approach uses statistical inference to find the best class for a given instance.

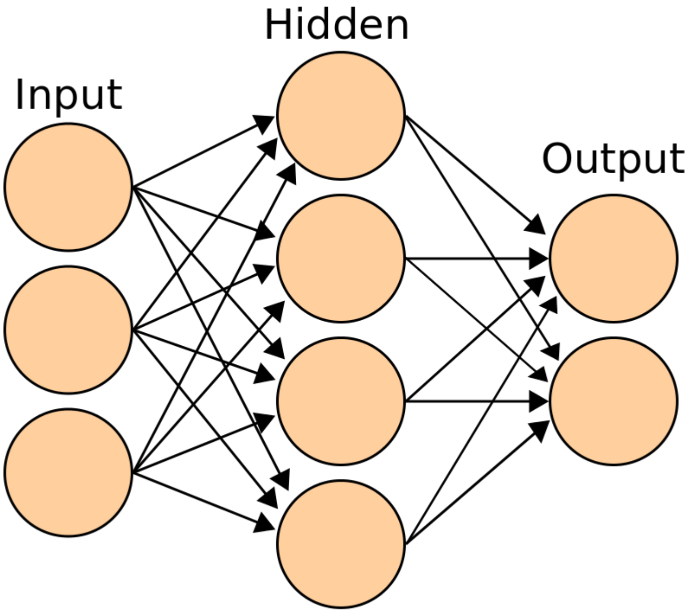


Figure 15: Neural network simple graphical structure

The most commonly used classification algorithms are:

- **Naive-Bayes classifier:** it makes use of simple 'probabilistic classifier' based on applying Bayes' theorem with Naive independence assumptions between features.

- **K-nearest neighbor:** k-NN classification has a class membership as output. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is typically a small integer value). If k=1, then the object is simply assigned to the class of that single nearest neighbor.

- **Decision Tree:** in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

- **Artificial Neural Network:** ANN is based on the connection of units or nodes called artificial neurons, which loosely remind the neurons in a biological brain. Each connection can transmit a signal to other neurons. An artificial neuron receives multiple signals as input and processes a single output that would become the input of other neurons. Neurons typically have a weight that adjusts as learning proceeds.

2.3.2 Regression problem

Regression algorithms belong to the family of Supervised Machine Learning. Regression algorithms predict the output values based on input features from the data fed in the system. The go-to methodology is the algorithm builds a model on the features of training data and using the model to predict the value for new data. Today, regression models have many applications, particularly in financial forecasting, trend analysis, marketing, time series prediction and even drug response modeling. Some of the popular types of regression algorithms are linear regression, polynomial regression, lasso regression and multivariate regression.

- **Simple Linear Regression model:** it is a statistical method that study relationships between two continuous (quantitative) variables. In linear regression, a model assumes a linear relationship between the input variables (x) and the single output variable (y). In this way the output can be computed from a linear combination of the input variables. When there is a single input variable, the method is called a simple linear regression. When there are multiple input variables, the procedure is referred as multiple linear regression. Sometimes this algorithm is affected by underfitting problem when a linear relationship is not enough to estimate the output.
- **Polynomial Regression model:** the main difference between this algorithm and the previous one is that the model is not linear, it is slower but has a greater accuracy. The underfitting problem is thus avoided, on the contrary an overfitting one can arise. The overfitting is due to an excessive adaptation to the training set with the loss of ability to correctly estimate new data.
- **Lasso Regression:** LASSO stands for Least Absolute Selection Shrinkage Operator. Shrinkage is defined as a constraint on parameters. Lasso regression is aimed to obtain the subset of predictors that minimize prediction error for a quantitative response variable. The algorithm starts imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward a zero, then variables with a zero-regression coefficient are excluded from the model. Variables with non-zero regression coefficients variables are strongly associated with the response variable. This lasso regression analysis is basically variable selection method and it helps analysts to determine which of the predictors are most important.
- **Multivariate Regression:** this algorithm is used when there is more than one predictor variable in a multivariate regression model, so it is implemented to predict the response variable for a set of explanatory variables.

3 Model

3.1 Data structure

In this section, a global view on data acquisition and database organization is provided. The ISI-Welding Company developed his own software (Welding Management System) able to control and supervise several aspects of the welding process. It is possible to upload large databases, to analyze the waveform of the main variables, to pull out a lot of useful indices, to set dressing parameters and splash detection ones and much more. This software is active during the welding process, its characteristic sample time is 4 ms. However, the data acquisition is performed by means of sensors connected to an $FPGA$ with a sample time of 1 ms. The control functions are carried out by the $FPGA$ for its faster sampling and the consequent ability to detect in a faster way the incoming disturbance. For example, when an abrupt change of slope in the voltage waveform is detected, the $FPGA$ decrease the input current. The WMS acquires the values from the $FPGA$ according to its sample time (so one acquisition every four $FPGA$ samples) and implements a data filtering, computes power and resistance from current and voltage, pull out min, max and mean of each parameter and so on.
As regard the database, the organization is developed on three levels:

$$\begin{gathered} Group \\ \Downarrow \\ Program \\ \Downarrow \\ Spot \end{gathered}$$

The group is the welding gun. The working station for the data collection is composed by two parallel operative lines each with 6 welding guns, so there are 12 groups. Then, a welding gun perform a series of welding points that can have different working conditions. A program is the set of all these conditions (i.e., number of metal sheet, welding time), every group can have different numbers of programs (usually $15 \div 20$). Finally, the spot is the welding point which current and voltage is measured. The welding guns can stop to perform the points defined by the set of programs if an alert situation is detected or for the electrode dressing, that is fixed after a certain number of points. Anyway, if this number of points is reached and the gun has still not finished the set of programs, the priority is given to correctly terminating the set and then the electrode is dressed. An important strategy developed by the company is the introduction of short-circuit points before and after the electrode dressing. These points can be useful to analyze the pollution of the electrodes and its effects on the electrical variables.
The most ambitious objective is to find a model and predictive algorithm that works in general, applicable to each welding gun regardless of the programs, but currently there are not enough information on the system to achieve this goal. Therefore, in this work the studies are carried out on welding spots that belong to the same program since, having the same working conditions, it is more probable to find common patterns on the variables.

3.2 ARX, ARMAX and OE

System identification is aimed at constructing or selecting mathematical models M by dynamical data, generated by a system S, to serve certain purpose (forecast, diagnostic, control, etc.).
The first step is to determine a class m of models within to search the most suitable model. There are 2 possible models to find:

- **transfer-function models**
- **state-space models**

The system identification problem may be solved using an iterative approach:

- Collect the data set:
  - design the experiment so that the data can be maximally informative.
  - pre-filtering technique of the data.
- Choose the model set or the model structure:
  - physical model with some unknown parameters may be constructed by exploiting the possible a-priori knowledge and insight.
  - black-box model may be employed, in this case the given data are elaborated without a physical reference.
  - gray-box model may be used, with adjustable parameters having physical interpretation.
- Determine the suitable complexity level of the model set or model structure.
- Tune the parameters to pick the 'best' model in the set (guided by data).
- Perform a model validation test.

At the end of this model development cycle, if the model approximates the behavior of the real data it is possible to use it, otherwise there is the necessity to restart from the beginning criticizing the data, the model orders or the other choices made in the development phase.
One of the approaches used to find a relationship between input and output is the polynomial identification. In this way we assume to have a completely unknown system (black-box) with only the measured data. This approach has been used in order to see if the polynomial relationship could suggest something about the physical model.

3.2.1 Polynomial model

Principal families of dynamic model can be considered as a particular case of:

$$y(t) = G(z)u(t) + H(z)e(t) \tag{1}$$

where $y(t)$ is the measured output, $u(t)$ is the command input and $e(t)$ is the error. $G(z)$ and $H(z)$ are transfer functions, given from the relationship between polynomials, in which parameters have to be estimated with precise identification methods.

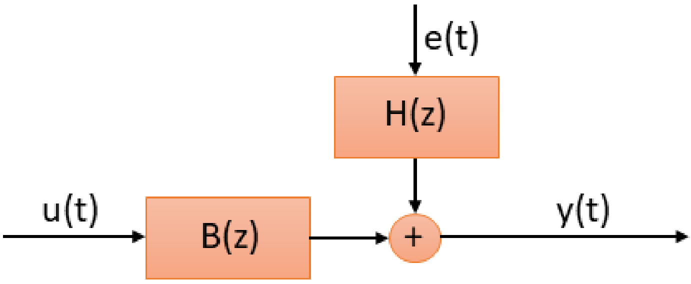


Figure 16: Example of the system to identify

The signal $v(t) = H(z)e(t)$ can be seen both as a noisy agent on the overall system or as a phenomenon not modeled from $G(z)u(t)$.
Moreover, it is possible to suppose that $H(z)$ has all the poles with modulus less than 1 and roots of the numerator with modulus less or at least equal to 1.
Specializing $G(z)$ and $H(z)$ in particular structures, it is possible to obtain different families of black-box model.

3.2.2 ARX model structure

An 'AutoRegressive eXogenous' model has the form:

$$y(t) = -a_1 y(t-1) - ... - a_{n_a} y(t-n_a) + b_1 u(t-1) + ... + b_{n_b} u(t-n_b) + e(t) \tag{2}$$

The noise enter as a direct error. If z^{-1} is denoted as the unitary delay operator such that $z^{-1}y(t) = y(t-1)$ and $z^{-2}y(t) = y(t-2)$, is possible to define:

$$\begin{aligned} A(z) &= 1 + a_1 z^{-1} + a_2 z^{-2} + ... + a_{n_a} z^{-n_a} \\ B(z) &= b_1 z^{-1} + b_2 z^{-2} + ... + b_{n_b} z^{-n_b} \end{aligned} \tag{3}$$

then, the above relationship can be written as:

$$A(z)y(t) = B(z)u(t) + e(t) \Rightarrow y(t) = \frac{B(z)}{A(z)}u(t) + \frac{1}{A(z)}e(t) = G(z)u(t) + H(z)e(t) \quad (4)$$

where:

$$G(z) = \frac{B(z)}{A(z)}, H(z) = \frac{1}{A(z)} \quad (5)$$

The block scheme is:

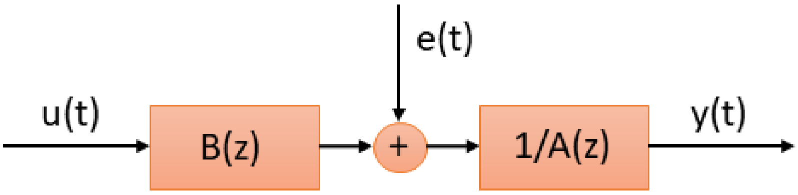


Figure 17: Example of an ARX model block diagram

It is possible to see how the noise pass through the term $1/A(z)$, the meaning is that the noise acts on the state of the system. The input is known as exogenous variable, then the model contains the **autoregressive (AR)** $A(z)$ and the **exogenous (X)** $B(z)$ parts. The integers n_a and n_b are the orders of these two parts of the ARX model.[7]

3.2.3 ARMAX model structure

The input-output relationship of the 'AutoRegressive Moving Average eXogenous' model is a difference linear equation:

$$\begin{gathered} y(t) + a_1 y(t-1) + a_2 y(t-2) + ... + a_{n_a} y(t - n_a) = \\ b_1 u(t-1) + ... + b_{n_b} u(t - n_b) + e(t) + c_1 e(t-1) + ... + c_{n_c} e(t - n_c) \end{gathered} \quad (6)$$

where the white-noise $e(t)$ enters as a linear combination of $n_c + 1$ samples.
By introducing the polynomials:

$$\begin{aligned} A(z) &= 1 + a_1 z^{-1} + ... + a_{n_a} z^{-n_a} \\ B(z) &= b_1 z^{-1} + ... + b_{n_b} z^{-n_b} \\ C(z) &= 1 + c_1 z^{-1} + ... + c_{n_c} z^{-n_c} \end{aligned} \quad (7)$$

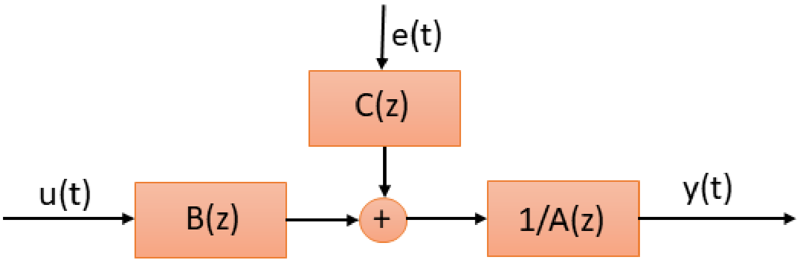


Figure 18: Example of an ARMAX model block diagram

The input output relationship can be written as:

$$\begin{aligned} A(z)y(t) &= B(z)u(t) + C(z)e(t) \\ y(t) &= \frac{B(z)}{A(z)}u(t) + \frac{C(z)}{A(z)}u(t) = G(z)u(t) + H(z)e(t) \end{aligned} \tag{8}$$

where: $G(z) = B(z)/A(z)$ and $H(z) = C(z)/A(z)$.
The **auto-regressive (AR)** part is included in the term $A(z)y(t)$, the **exogenous (X)** part in $B(z)u(t)$ and the **moving average (MA)** part in $C(z)e(t)$ (which is a colored noise instead of the white one).
The integers n_a,n_b,n_c are the orders of these three parts of the ARMAX model (ARMAX(n_a, n_b, n_c)).[7]

3.2.4 OE model structure

The relationship between input and undisturbed output is a linear difference equation:

$$w(t) + f_1 w(t-1) + ... + f_{n_f} w(t - nf) = b_1 u(t-1) + ... + b_{n_b} u(t - n_b) \tag{9}$$

and the model output is corrupted by white measurement noise:

$$y(t) = w(t) + e(t) \tag{10}$$

By introducing the polynomials:

$$F(z) = 1 + f_1 z^(-1) + ... + f_{n_f} z(-n_f) \tag{11}$$

$$B(z) = b_1 z^{-1} + b_2 z^{-2} + ... + b_{n_b} z^{-n_b} \tag{12}$$

The above input-undisturbed output relationship can be written as:

$$\begin{gathered} F(z)w(t) = B(z)u(t) \Rightarrow \\ y(t) = w(t){+}e(t) = \frac{B(z)}{F(z)}u(t) + e(t) = \\ G(z)u(t) + e(t) \end{gathered} \tag{13}$$

where $G(z) = B(z)/F(z)$.
The integers n_b and n_f are the orders of the OE model, denoted as OE(n_b,n_f).

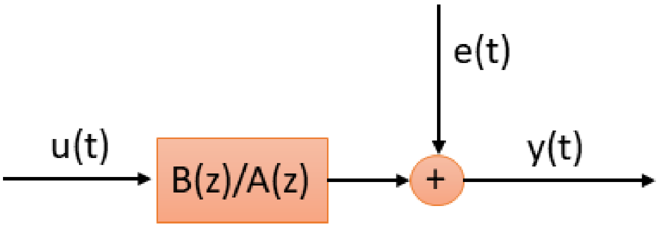


Figure 19: Example of an OE model block diagram

In the ARX and ARMAX model the poly A(z) is the denominator of every component, and this is often an uncomfortable situation (too much restrictive).
To relax this hypothesis, it is possible to use the OE structure in order to have a better simulator of the real plant.
On the other hand, if a one-step predictor of the system is needed, the ARX structure gives better results.[7]

3.2.5 Data processing

The data measured during the short circuit welding point has been used to search the polynomial model.
As a matter of fact, this kind of welding point are taken from the company before and after the dressing of the electrodes to take track of resistance changes during the dressing.
In particular, the short circuit welding point considered are the ones taken after the dressing. This to try to identify the most ideal condition possible, without the uncertainty introduced by the electrodes worn out and the metal sheets interposed.
The data have been filtered with a low pass filter, in order to reduce the noise and the mean value has been removed.[8]

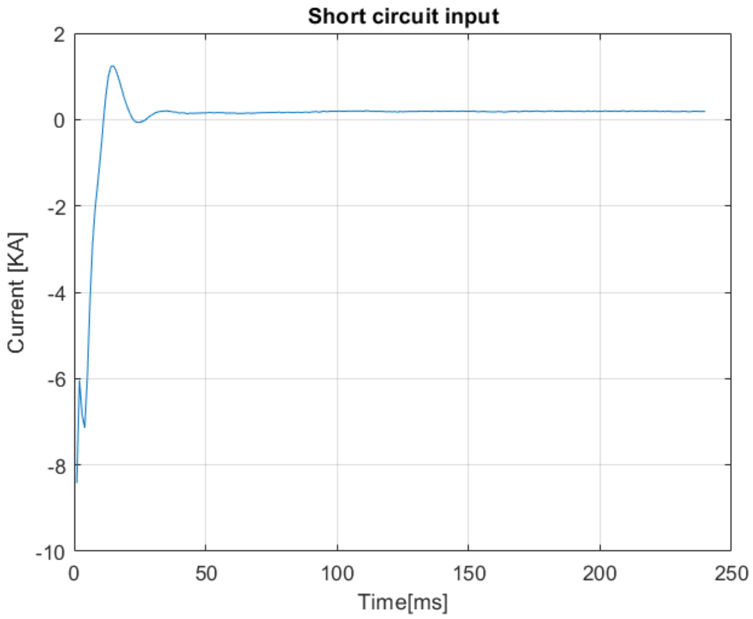


Figure 20: Example input in a short circuit welding point

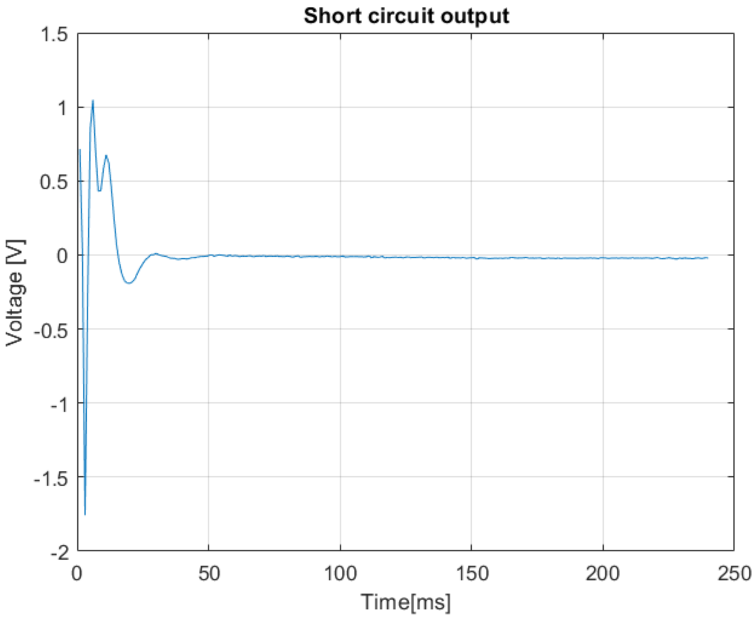


Figure 21: Example output in a short circuit welding point

It is possible to see that the welding point last $240ms$.
The first $40ms$ are considered 'blanking time' because they are really noisy and, possibly, they don't bring helpful information. For this reason, in the evaluation of the model fitting, they will be not considered.

3.2.6 Results

With the selected data, different kind of ARX, ARMAX and OE models have been tried. The M set of possible models has been chosen always with a possible delay from 1 to 5, and every single component of the order from 1 to 5 too.

Model	Order	Delay
ARX	$n_a = n_b = 1$ to 5	$n_k =$1 to 5
ARMAX	$n_a = n_b = n_c = 1$ to 5	$n_k =$1 to 5
OE	$n_b = n_f = 1$ to 5	$n_k =$1 to 5

Table 3: ARX, ARMAX and OE orders and delays

n_k is called delay because is the first useful time instant of the input so that the equation, for example in the ARX model become:

$$y(t) = -a_1 y(t-1) - ... - a_{n_a} y(t-n_a) + b_1 u(t-n_k) + ... + b_{n_b} u(t-n_b-nk+1) + e(t) \quad (14)$$

and the same for ARMAX and OE models.
The best models are firstly evaluated according to 99% auto-correlation region.

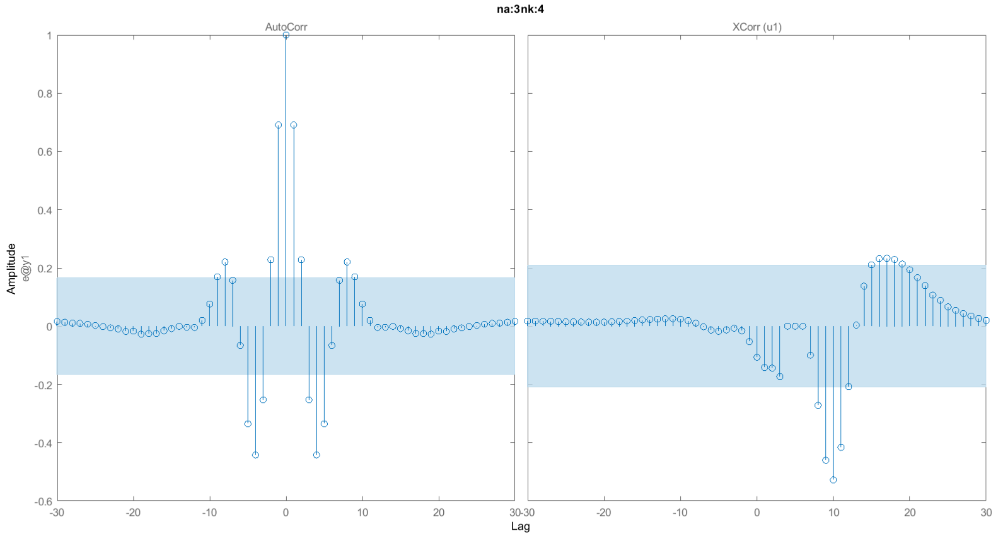


Figure 22: Auto-correlation and cross-correlation analysis

The choice of the best models is made evaluating the number of residues out of the 99% confidence zone in the right half plane (due to plot symmetry) of the auto-correlation.

In the Figure 22 it is possible to visualize the auto-correlation and cross-correlation function of the ARX model with $n_a = 3$, $n_b = 3$ and $n_k = 4$. According to the results, a model is evaluated if their auto-correlation plot shows a maximum of three residues out of the 99% confidence region, otherwise the model is not acceptable.

Furthermore, two input/output choices have been considered:

- current input and voltage output,
- current and squared current inputs and voltage output.

As example, below are reported some of the best ARX models with double input and their auto-correlation:

n_a	n_k	Resid
1	1	1
3	1	1
2	5	1
5	4	1
4	4	2
5	3	1
4	3	1
3	2	2
5	5	1

Table 4: ARX residues analysis results

Resid stands for the number of residues outside the 99% confidence region, while n_a is the reference for the order and n_k is the reference for the delay.
All the model reported are good choice to polynomial identification, the best of them has been selected with Root Mean Square Error analysis.
The Mean Squared Error is defined as:

$$\mathbf{MSE} = \frac{1}{N - N_0} \sum_{N_0+1}^{N} (y(t) - \hat{y}(t, \theta))^2 \tag{15}$$

where N are the samples of the output, N_0 is the first sample taken into account for evaluation, $\bar{y}$ is the mean of the output values $y(t)$ and the Mean Squared Error (MSE). The MSE is a measure of the quality of the estimator, the closer its value is to zero, the better the estimator is. In fact, the MSE takes into account the variance of the estimator and the its bias with respect to the real output. However, it cannot be considered as a reliable index since it has the squared measurement unit of the output and it strongly depends on numerical values of data. Therefore a dimensionless fit parameter has to be chosen as the best index for performance evaluation. The RMSE is the square root of the MSE:

$$\textbf{RMSE} = \sqrt{MSE} \tag{16}$$

The RMSE represents the square root of the second sample moment of the differences between predicted values and observed values or the quadratic mean of these differences, it aggregates the magnitudes of the errors in predictions for various data points into a single measure of predictive power. In the following plot is possible to see how, basically, increasing the order, the RMSE tends to decrease. The goal is to choose the lower RMSE between the models that have been considered after the residual analysis.

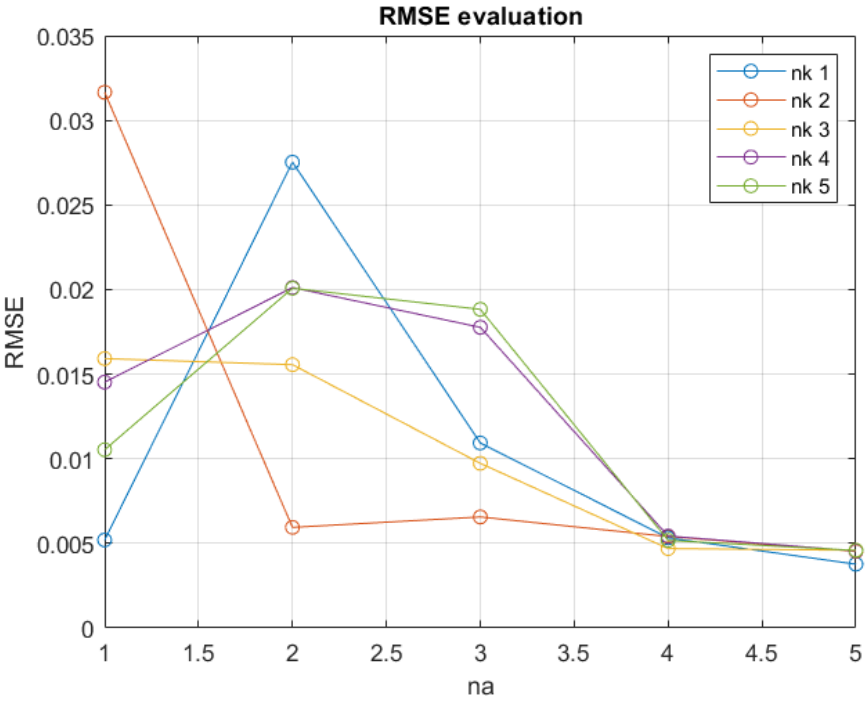


Figure 23: RMSE analysis of ARX model

It is possible to see that the ARX model with $n_a = 5$ and $n_k = 5$ has the lower possible RMSE (0.0026). For this reason the ARX(5,5,5) (ARX(n_a,n_b,n_k)) is the best possible model in the ARX set.

Since also the RMSE is not dimensionless, but it has the same dimension of the output, it occurs another parameter to estimate the quality of the estimation. The Best Fit parameter has been chosen for this purpose:

$$\textbf{Best Fit} = 1 - \sqrt{\frac{MSE}{\frac{1}{N-N_0}\sum_{t=N_0+1}^{N}(y(t)-\hat{y})^2}} \tag{17}$$

It can be easily converted in percentage.

At the end of this process, the result is a polynomial model able to produce the same output of the system if it is fed with the input. This model is just a mathematical representation and has not a physical meaning.

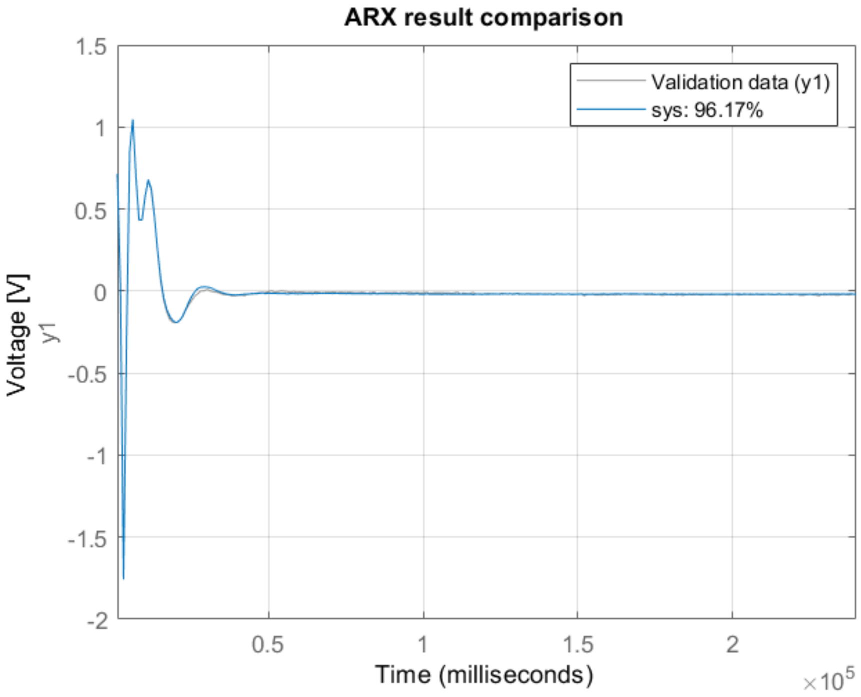


Figure 24: Measured and estimated output comparison

The same procedure can be iterated for the ARMAX and the OE model.
The results are:

Model	C	C and C^2
ARX	$n_a = n_b = 5$ and $nk = 2$	$n_a = n_b = 5$ and $nk = 5$
ARMAX	$n_a = n_b = n_c = 4$ and $n_k = 3$	$n_a = n_b = n_c = 5$ and $n_k = 3$
OE	$n_b = n_f = 4$ and $n_k = 3$	$n_b = n_f = 4$ and $n_k = 3$

Table 5: ARX, ARMAX and OE orders and delays for both input conditions

One thing to specify is that in the situation with current and its squared value as input the coefficient relatives to the inputs have to be doubled.
The Multiple Input Single Output (MISO) system has to be considered as the superposition of two Single Input Single Output (SISO) systems.
To report an example, the ARMAX MISO system will be of the type ARMAX$(n_a, n_{b1}, n_{b2}, n_c, n_{k1}, n_{k2})$ where n_{b1} and n_{k1} are related to the first input (C) while n_{b2} and n_{k2} to the second one (C^2). As simplification, $n_{b1} = n_{b2}$ and $n_{k1} = n_{k2}$ have always been considered.
The obtained results are encouraging.
The data set was previously divided in 70% for the estimation of the model and a 30% for the validation.
The results during validation show a floating fitting between 80% and 95%.

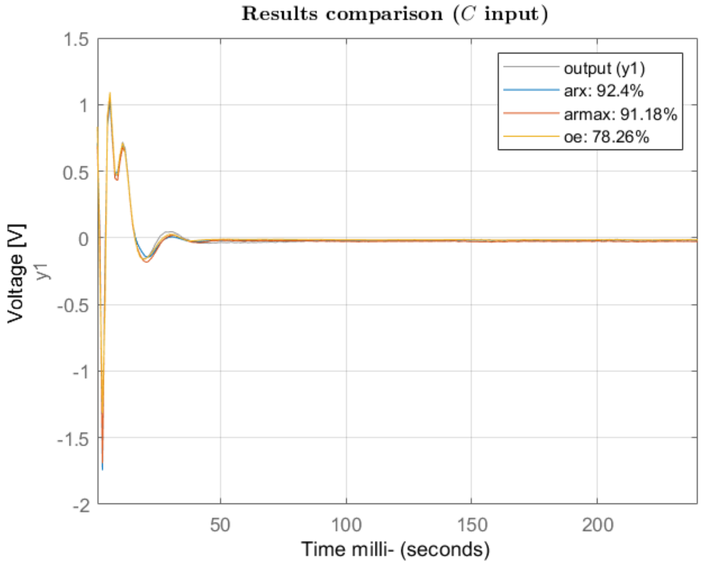


Figure 25: Measured and estimated output (single input models)

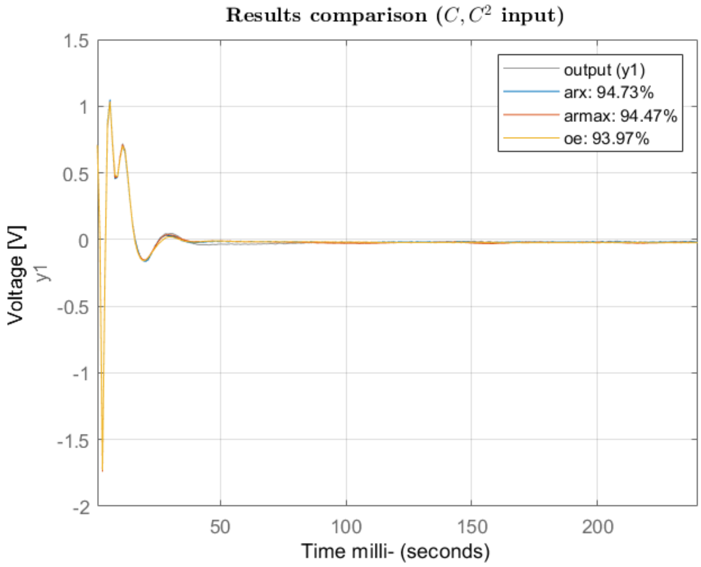


Figure 26: Measured and estimated output (double input models)

It is possible to see that, in general, the results obtained with the double input (C, C^2) are better than the single input cases.
This shows a possible physical relationship between the output voltage and the power of the system, proportional to the squared current $(P = I^2R)$.
Another attempt was also made considering as input also the cube value of the current. In this case the fitting percentage did not increase significantly as the C^2 case.
A possible meaning of this phenomenon is that the cube of the current has no physical meaning.

3.3 NARX Model

The nonlinear autoregressive network with exogenous inputs (NARX) is a recurrent dynamic network with feedback connections enclosing several layers of the network. This particular application is very useful with time-series data. It can be used as a predictor, for nonlinear filtering and for the modeling of nonlinear dynamic systems. The defining equation of the NARX model is:

$$y(t) = f(y(t-1), ..., y(t-n_y), u(t-1), ..., u(t-n_u)) \tag{18}$$

where the next value of the dependent output signal $y(t)$ is regressed on previous values of the output signal and previous values of the independent input signal. It is possible to design a feedforward neural network to approximate the function f. So, the output of the network is an estimate of the output of the examined system.[9]

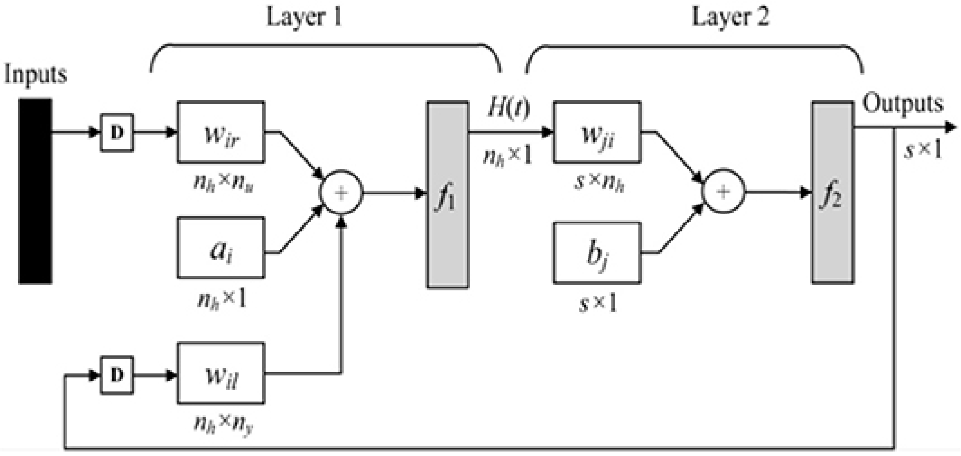


Figure 27: NARX architecture

During the training, it is more convenient to use the true output as delayed input instead of the estimated one, in an open-loop architecture. In this way the inputs of the feedforward network are more accurate and the neural network is better trained.

After the training, there are two different possible ways to implement this algorithm according on the available measurements. If the output is measured during the process, it is possible to feed the network with its real values, without feeding back the estimated ones. On the other hand, if the output is not measured or its measures are not readily available, it is possible to close the loop and to feed back the estimated output as new network input.

Different types of neural network architecture have been tested, changing input delays and layers size. The performance has been evaluated through the best fit parameter.

3.3.1 Results

Different tests have been performed using the NARX methodology. At first, short-circuit data have been used, in particular short-circuit data after the electrode dressing. This data represents the most 'ideal' conditions that it is possible to obtain on welding guns, without disturbances introduced by metal sheets and electrode pollution, so they are useful to build and evaluate a model.

Against this background, a neural network with 10 delayed states (both external input and feedback output) and two hidden layers with 30 and 5 neurons has been implemented. A training set of 70 welding spot data has been fed to the network. A first test is made with the open-loop network. This is a realistic choice because the input current and the output voltage are measured and readily available in order to detect disturbances. In addition, it is clear that the network has a higher accuracy feeding back real values than the estimated ones.

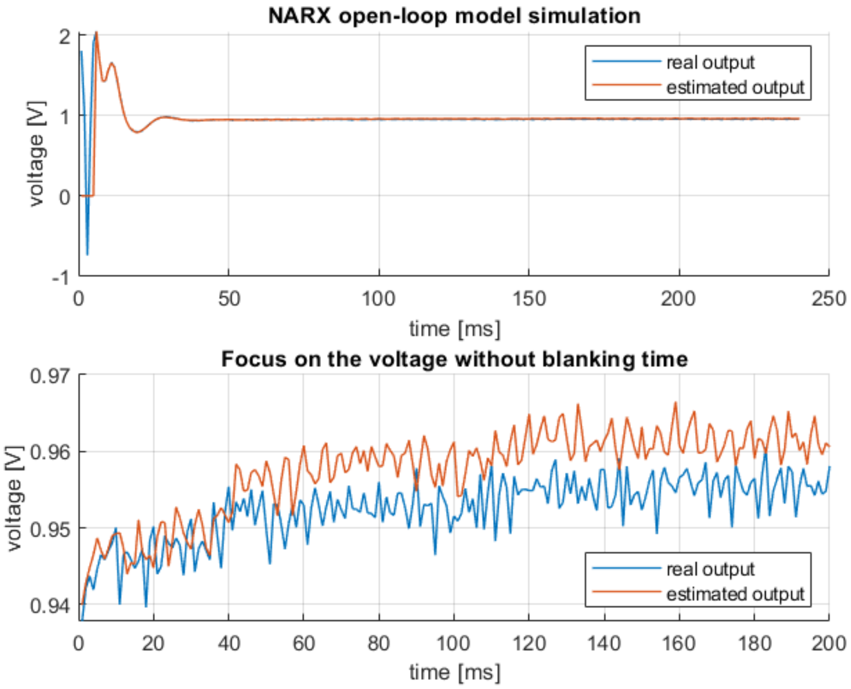


Figure 28: Comparison between the real and open-loop estimated output

In this simulation, the MSE is $3.73 \cdot 10^{-5}$ while the fit is about 95.2%. The simulation starts with a 10 ms delay, this is due to the initial acquisition of the delayed input and output by the neural network that can begin to estimate only once acquired these data.

An interesting application of this algorithm is obtained by closing the loop of the neural network. This can be a choice when the output is not available while the system is working and the neural network takes his own estimated output as input for the estimation of the subsequent step. Unfortunately, in this case the simulation does not lead to good results:

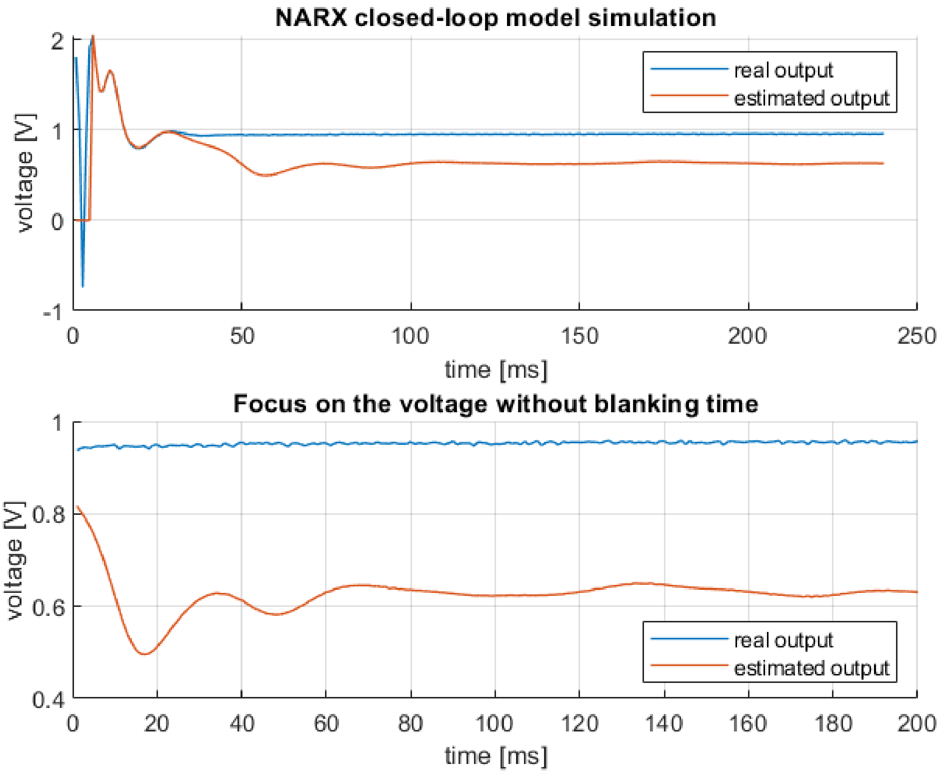


Figure 29: Comparison between the real and closed-loop estimated output

It is clear that the estimated output is not able to approximate the real one. This 'failure' may be due to the incapability of the electrical data to fully describe the phenomena involved in a welding process.

Anyway, it can also be considered that the most significant data are the ones collected after the blanking time (first 40 ms). A better estimation with the closed-loop is obtained by removing both in training and test data those values collected during the blanking time and re-designing the neural network for the new settings. The following figure shows the result with 5 delayed states and two hidden layers with 20 and 10 neurons: the estimation is still not excellent, but it is able to follow the real output also in closed-loop, without any information on its real values.

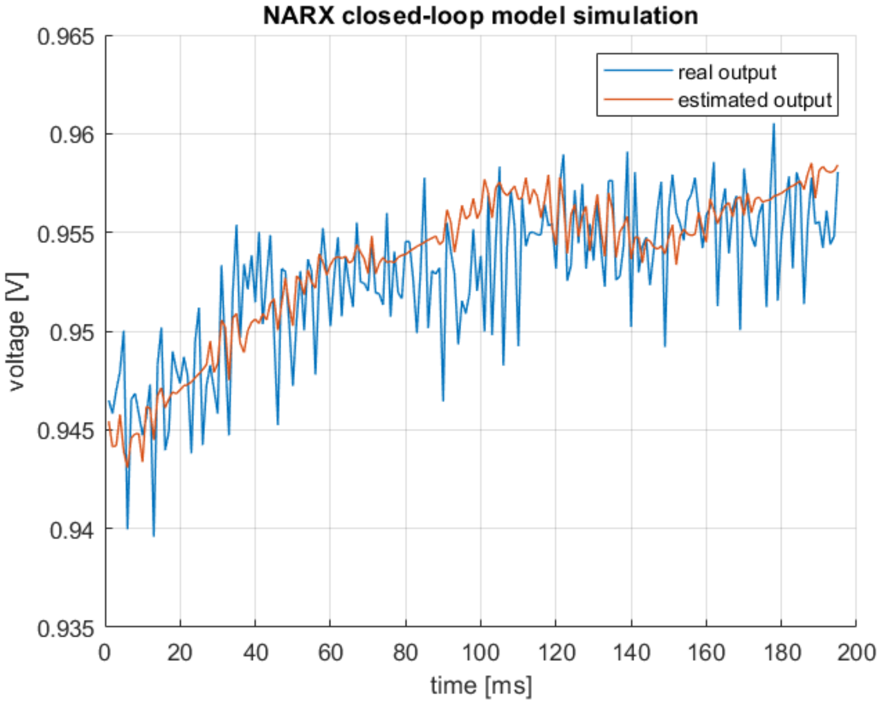


Figure 30: Comparison between the real and closed-loop estimated output without blanking

Another attempt is made using data collected from short-circuit points before the electrode dressing, after 100 real welding points. In this case the disturbances introduced by metal sheets are still avoided, but the electrode is polluted. A neural network with 10 delayed states and two respectively 40 and 10 neurons hidden layers has been designed and trained with the new dataset. Performance are similar to the previous case, with a MSE of $2.79 \cdot 10^{-5}$ and a fit of about 93%:

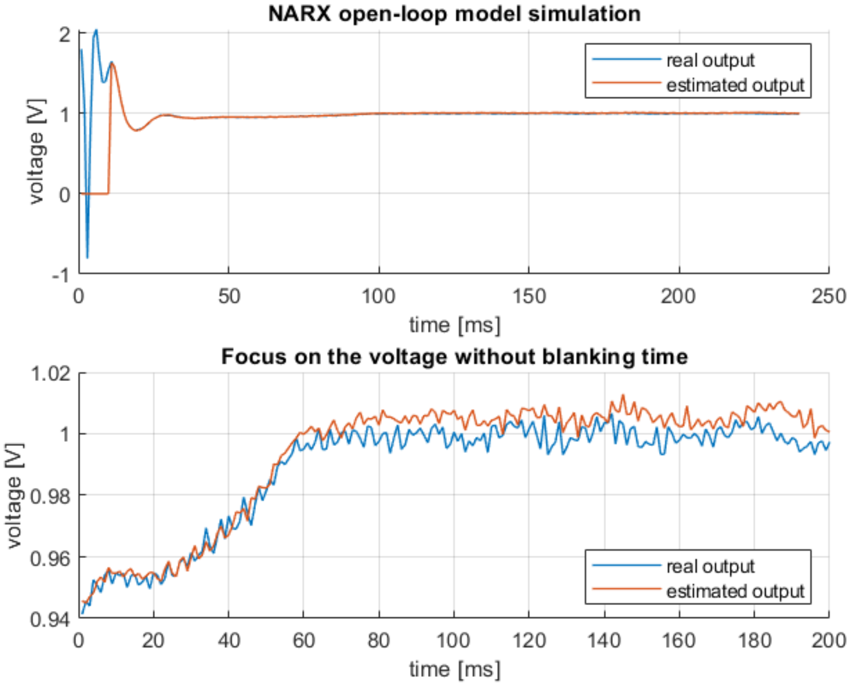


Figure 31: Comparison between the real and open-loop estimated output

The last possible attempt is made using real welding data, considering disturbances and discovering if these data can be adequate for a model. 10 delayed states, a 30 and a 5 neurons hidden layers are the characteristics of the neural network. There is higher availability of real welding points with respect to the short-circuit ones, in fact a set of 500 points has been selected as training set. There are points affected by splash. A first test has been performed on a welding point that has not presented the splash: with a MSE of $3.1 \cdot 10^{-5}$ and a 95.4% fit, the estimation is able to approximate the real output.

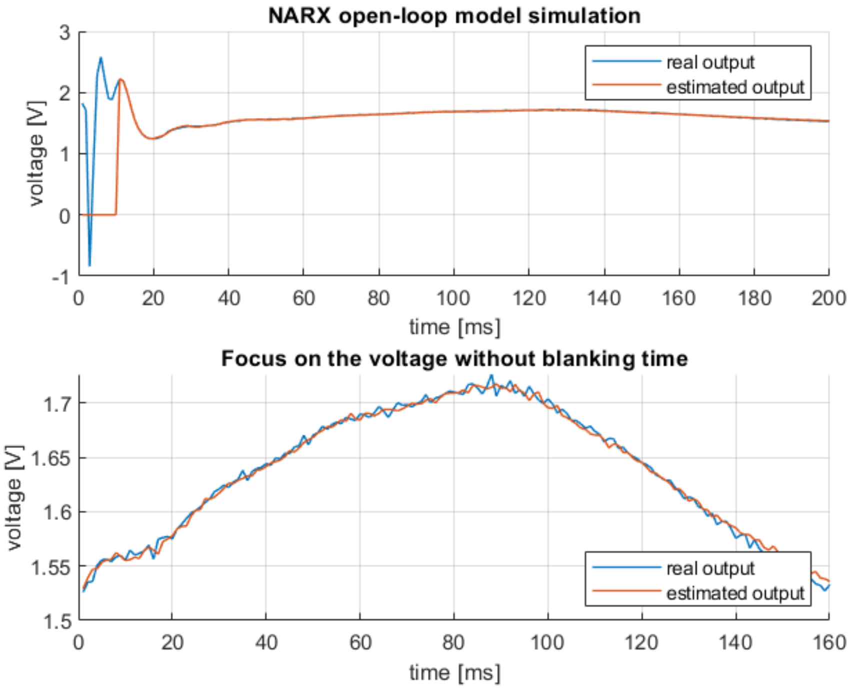


Figure 32: Comparison between the real and open-loop estimated output, without splash

The following figure represents the test implemented using a point affected by splash:

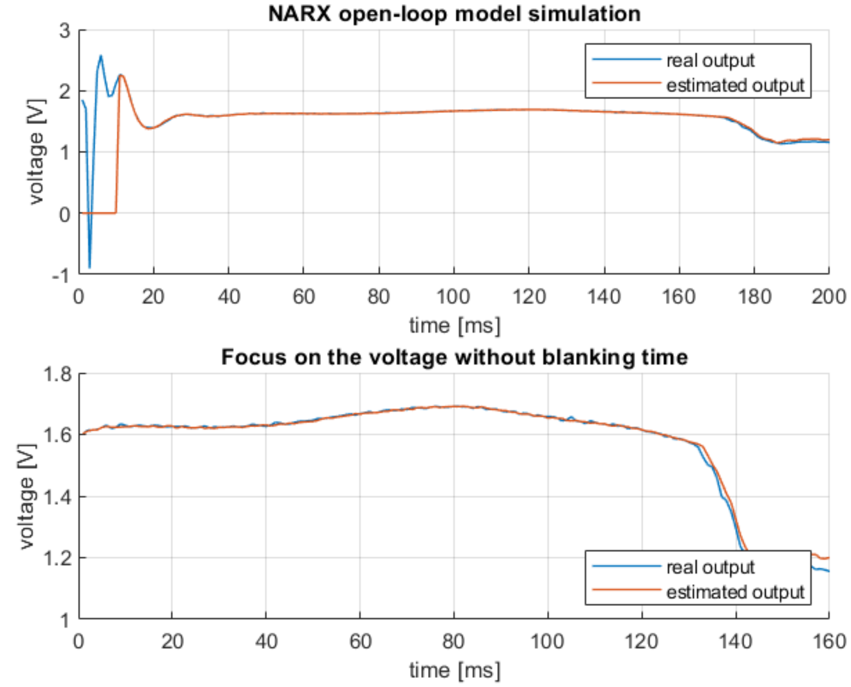


Figure 33: Comparison between the real and open-loop estimated output, with splash

The indices are worst (MSE=$1.86 \cdot 10^{-4}$, fit=92%) but the general trend is still acceptable. Actually, the splash is well estimated and this open the way to a possible forecasting attempt using the same theoretical basis of this algorithm, with the necessary adaptation for the different kind of problem. A final closed-loop test is performed:

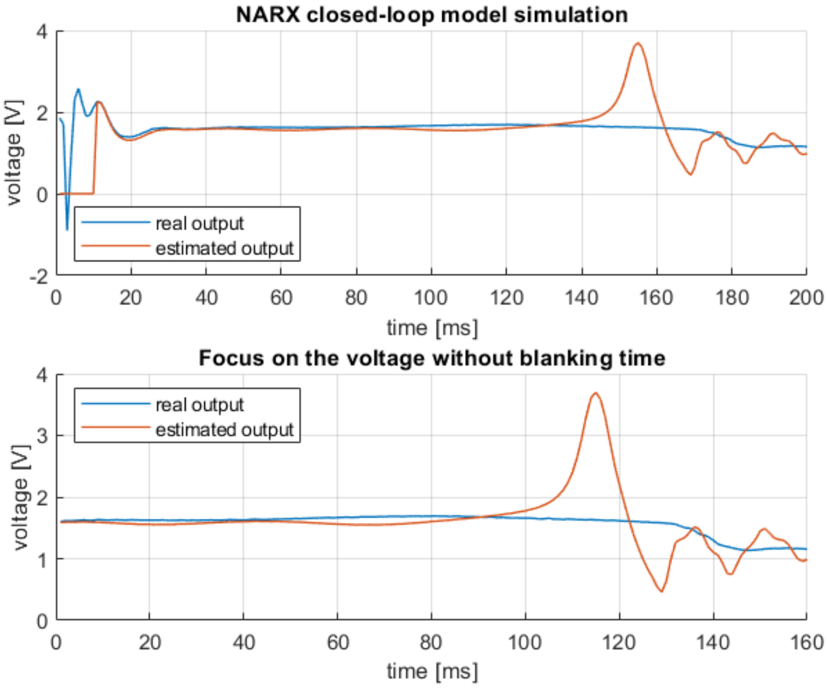


Figure 34: Comparison between the real and closed-loop estimated output, with splash

As expected, the estimation is completely inappropriate. There are too many factors not considered and not yet available that affect the welding process.

3.4 Physical model

From the physical point of view, the main role in the welding process is played by the Joule effect. This phenomenon is defined as the process by which the passage of an electric current through a conductor produces heat. The power of heating generated by an electrical conductor is proportional to the product of its resistance and the square of the current. When two metal sheets are put in touch and compressed, the highest electric resistance can be found at the point of contact between the sheets. So, during the welding process, the higher power of heating is released in this point and here the metal starts to melt.

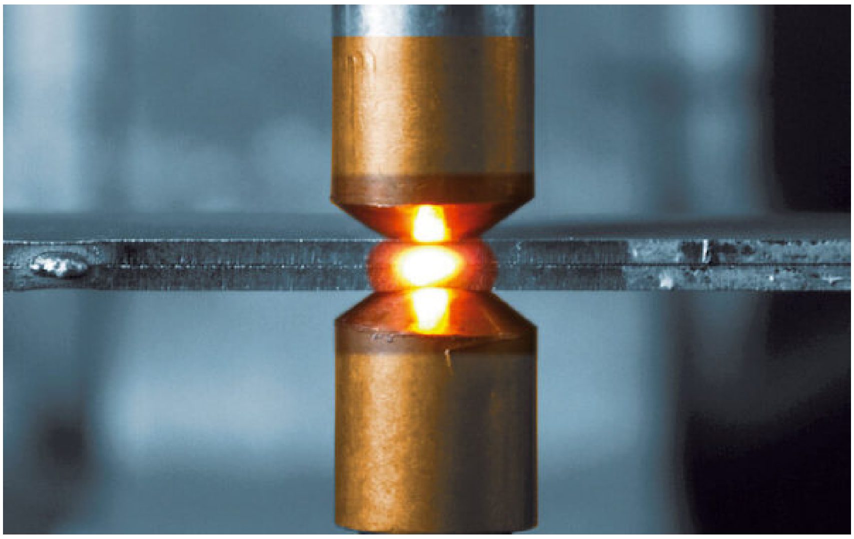


Figure 35: Weld core of metal sheets during a welding process

It is important that the welding process last long enough to allow an effective mixing of the melted metal of the sheets. In fact, a brief release of current can lead to a 'gluing' of the sheets rather than a welding. On the other hand, a longer one lead to a possible expansion of the welding core out of the contact zone with the electrodes, causing splash or a less controlled welding. Those reasons make explicit the importance of the power involved during the process. The ISI-Welding Company evaluates the power as an index of the quality of the point and applies also a control based on the released power during the welding process: if the energy is too low, it makes the process last longer; if it is too high, it makes it stops prematurely.
The dependence of the welding process by the power was also suggested by the previous studies on the polynomial models since they perform better with the information on the square of the current (directly proportional to the power). In general, the knowledge on the spot welding states that the resistance spot welding has to be seen as an **electric**, **mechanic** and **thermal** phenomenon at the same time.[10]

- **Electrical conductivity:** current, necessary to generate a certain quantity of heat, increases with increasing conductivity.

- **Thermal conductivity:** current necessary to compensate the heat dissipated by conduction increases with increasing thermal conductivity.
- **Thermal expansion coefficient:** higher current value and shorter time are needed to avoid the expulsion of the melted core and guarantee sufficient heat at the same time.[11]

In the last point it emerges that electrical quantities are also greatly influenced from the metal fusion and its consequent cooling. For those reasons, physically, there is the necessity to investigate the dependency between the current flowing into the circuit, energy variation and resistance variation. Unfortunately, power and resistance data are not measured and it is hard to identify a precise data-driven model.

The system can be classified as a **grey-box**: this is an approach that combines a partial theoretical structure with data to complete the model. Thus, grey box models are an hybrid between black box where no model form is assumed and white box that are purely theoretical.

It is known that the voltage depends on the current, its derivative for inductive phenomena and its integer value for capacitive phenomena. Moreover, it depends also on the power and its derivative (to take into account the thermal effects), on the resistance and its derivative (to take into account the thermal expansion and its change of state).

Once the theoretical basis have been laid and a large dataset is made available, a data-driven methodology has to be chosen to implement the physical model. The method of Least Squares is a standard approach in regression analysis to approximate the solution of overdetermined systems (sets of equations in which there are more equations than unknowns) by minimizing the sum of the squares of the residuals made in the results of every single equation. A residual is defined as the difference between the actual value of the dependent variable and the value predicted by the model:

$$r_i = y_i - f(x_i, \theta) \tag{19}$$

The most important application is in data fitting. Least-squares problems can belong to two categories: linear and nonlinear least squares, depending on the linearity of the residuals. The linear least-squares problem occurs in statistical regression analysis, the nonlinear problem is usually solved by iterative refinement in which at each iteration the system is approximated by a linear one, and the calculation is similar in both cases.

The only input available is the current. A first basic model can be designed considering the voltage as dependent only on the current and its square value, for the proportionality with the power. This equation represents the mathematical model:

$$V(t) = \theta_1 I(t) + \theta_2 I(t)^2 \tag{20}$$

This equation can be arranged in matrix form:

$$y = X\theta \tag{21}$$

Where y is an $n - by - 1$ vector of responses (voltage values), X is the $n - by - m$ design matrix for the model (current and its square), θ is the $m - by - 1$ vector of parameters to identify. The least square method makes use of the pseudo-inverse of a matrix:

$$A^+ = A^T(A^TA)^{-1} \tag{22}$$

This formula is a generalization of the inverse of a matrix in case it is not squared. Taking up the matrix equation of the model, with some mathematical tricks and using the pseudo-inverse the following steps can be performed:

$$\begin{gathered} X^TX\theta = X^Ty \\ \downarrow \\ \theta = (X^TX)^{-1}X^Ty \end{gathered} \tag{23}$$

In this way it is possible to compute θ from the collected inputs-outputs.
The first chosen dataset is the one with short-circuit point data. This is the simplest available, it is possible to obtain the first useful information from this model and, at a later stage, other complex models that take into account further factors can be implemented and checked. The matrix of parameters θ is therefore computed using the input-output set of selected data, then a test on a point is performed:

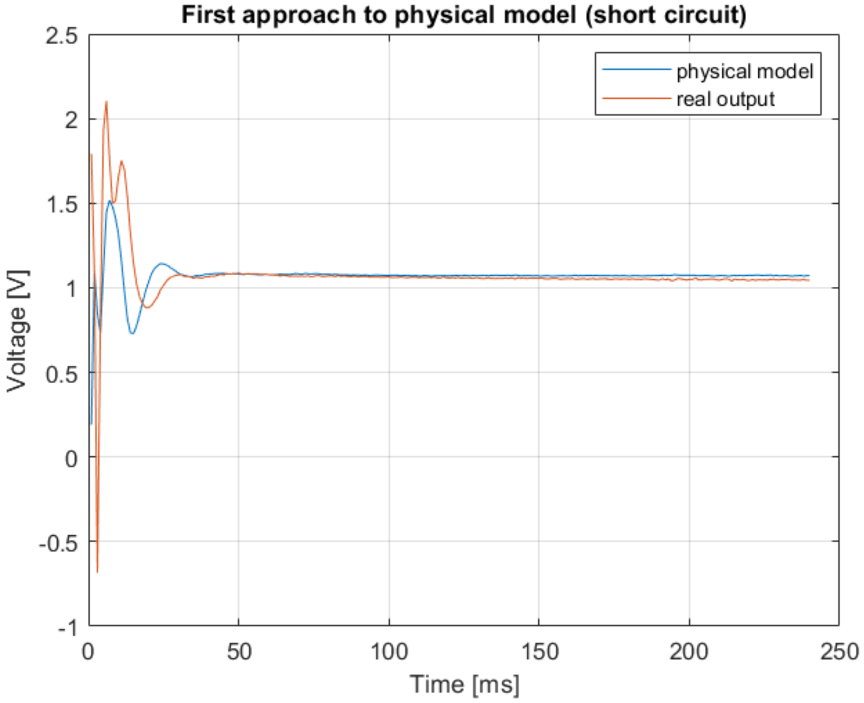


Figure 36: Comparison between estimated and real model (short circuit case)

Results show that the estimated model is not very able to approximate the real one for the entire duration of the signal.
Another attempt with this simple model is performed using a dataset of real welding spots.

The new parameters are computed, and another test is executed:

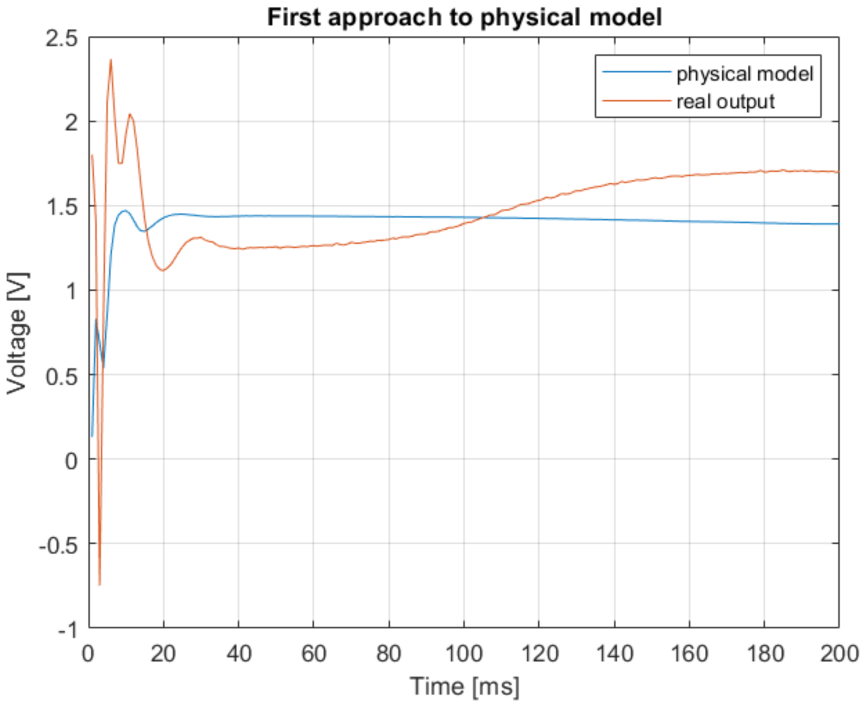


Figure 37: Comparison between estimated and real model (real welding case)

As expected, in this case the estimation is not acceptable. There is a simple reason: a lot of factors have not been intentionally considered, such as thermal and conductivity phenomena, the change of resistance due to the melting of the metal, capacitive and inductive phenomena that can arise in an electric circuit. All these factors in some way depends on the current. The following equation collects various shapes of the current which can be able to take in account these phenomena that are not directly measured. According to the Least Square method, every variable has been associated to a θ parameter.

$$V(t) = \theta_1 I(t) + \theta_2 I(t)^2 + \theta_3 \frac{dI(t)}{dt}\frac{1}{I(t)^2} + \theta_4 \frac{dI(t)}{dt} + \theta_5 \int I(t)dt \tag{24}$$

Moving over to the matrix, the equation does not change with respect to the previous one, what changes are the X and θ dimensions and the computational load will be higher. Of course, real welding points data are the best choice for this kind of model built on assumptions that implies the presence of metal sheets during the welding process. Subsequently, θ parameters have been computed and analyzed. The parameter associated to the integral of the current is very small and can be neglected. This can mean that there are no capacitive phenomena involved in the welding process. As last step, a test is executed:

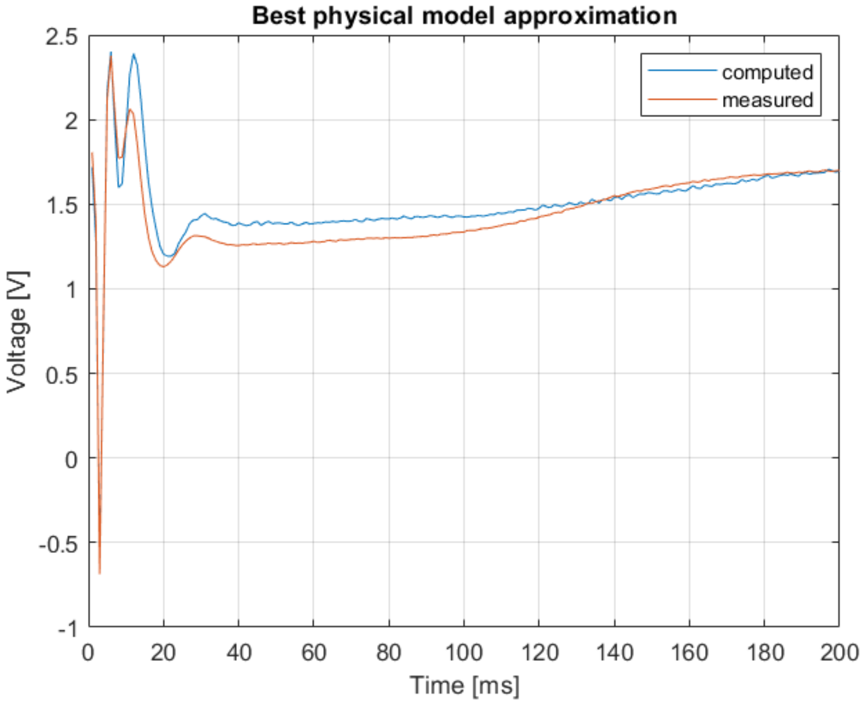


Figure 38: Comparison between physical model and real outputs

Results shows that the estimated model is able to roughly approximate the real output. Better performance can be achieved by measuring other important variables (that can be compression force, temperature, etc) and including them in the model.

4 Splash definition

With greater awareness of the phenomena involved in the welding process and with greater knowledge of the model, it is possible to proceed towards the development of a **predictive maintenance** algorithm.
First of all, it is essential to thoroughly understand the splash phenomenon. As already stated, it is defined as a spillage of melted metal out of the weld core:

Figure 39: Spillage of melted metal

There are many reasons why a splash can occur:

- lack of perpendicularity between the electrode axis and the sheets plane,
- weak pressure exerted by electrodes on sheets,
- electrodes pollution,
- excessive energy involved during the welding process.

The ejection of material between the electrodes lead to a decreasing of resistance of the electrical circuit. It is thus possible to detect analytically the presence of the splash by measuring the voltage: a decrease in resistance results in a decrease in voltage. The following figures sum up the difference of the measured variables between an optimal welding point and another one affected by splash.

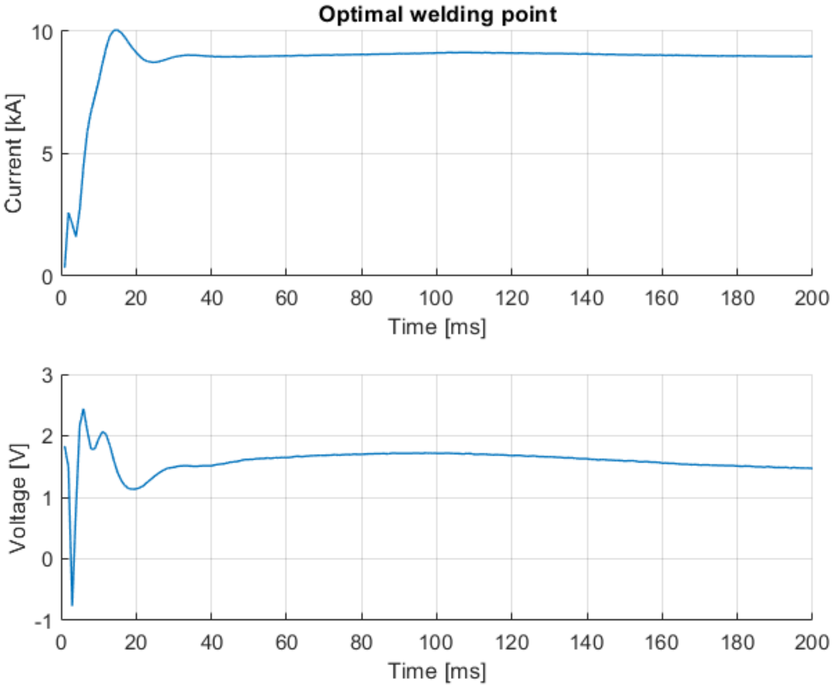


Figure 40: Optimal welding point

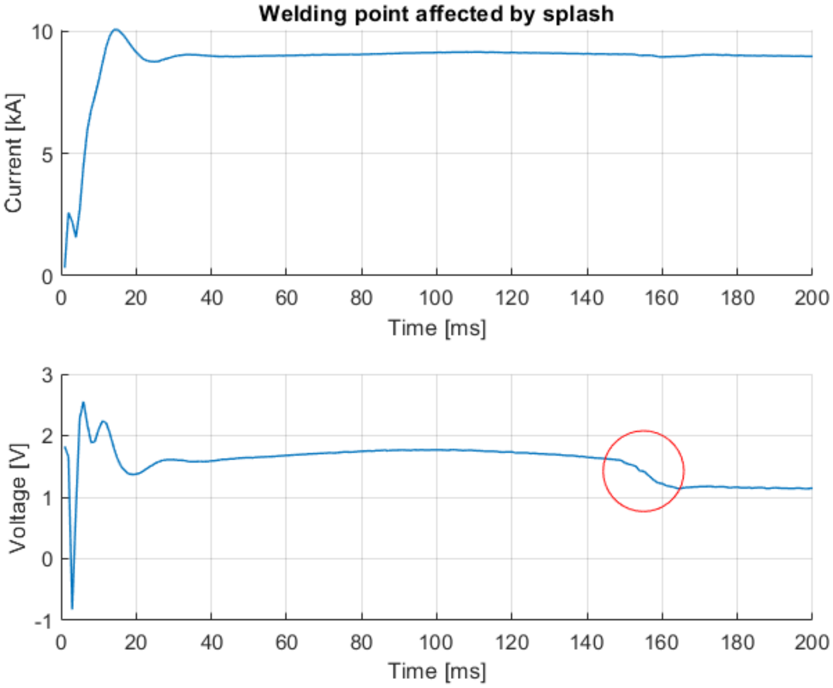


Figure 41: Welding point affected by splash

4.1 Splash classification

The presence of splash is therefore detectable, while it is occurring or once it has occurred. The present strategy of the company to limit the damage is to decrease the supplied current as soon as an abrupt decrease of voltage slope is detected. The real challenge is to find out some clues that allow the identification of an incoming splash from the data available.

Before focusing on this goal, it is important to understand if machine learning algorithms are able to classify the presence of splash after it showed up, at the end of the welding process. This information will not be useful for prediction purposes in a practical way, but it is useful to understand which machine learning algorithm to implement for the prediction. It is evident that an algorithm that is not able to classify a splash after its manifestation, will not be able to predict it.

As specified, welding points dataset can be classified on the basis of the group and the program to which they belong. Analyzing the database, it is possible to notice that there are programs that contain very few splashes compared to the number of total points. It is convenient to choose programs with a larger number of splashes, both because they are the programs on which it will be preferable to avoid this disturbance and to have a good number of examples to adequately train the algorithm.

The chosen training dataset (available in the database in PLD, group 02, program 005) is made up of 1600 points, 1385 without splash and 215 with splash. Each welding process lasts 200 ms, the acquisition sample time of the $FPGA$ is 1 ms so there are 200 input and output values for each point. As regard the machine learning algorithm, the artificial neural networks are a good choice to implement a classification. A first artificial neural network is therefore designed, with a 200 input layers and a single output layer. The input corresponds to the measured values of the voltage, the output is set to be 0 if the point is not affected by splash and 1 if affected (safe and warning condition respectively). Two hidden layers (50 and 5) are also designed.

The dataset has been divided in 1200 points for training set and 400 for testing. The algorithm uses the scaled conjugate backpropagation gradient to improve the network performance. This performance is evaluated through the cross-entropy function, defined as:

$$CrossEntropy = -t * log(y) \tag{25}$$

in which t is the target (or true output) and y the network output. The function returns a result that heavily penalizes outputs that are extremely inaccurate (y near 1-t), with very little penalty for fairly correct classifications (y near t). Minimizing cross-entropy leads to good classifiers. The aggregate cross-entropy performance is the mean of the individual values. The initial training set is subdivided in 70% training set and 30% validation set. Summing up, the training set is used to compute the layers parameters and the backpropagation gradient algorithm is applied on this set, then the performance of these parameters is evaluated on the validation set and when the optimal cross-entropy value is detected, the train is stopped:

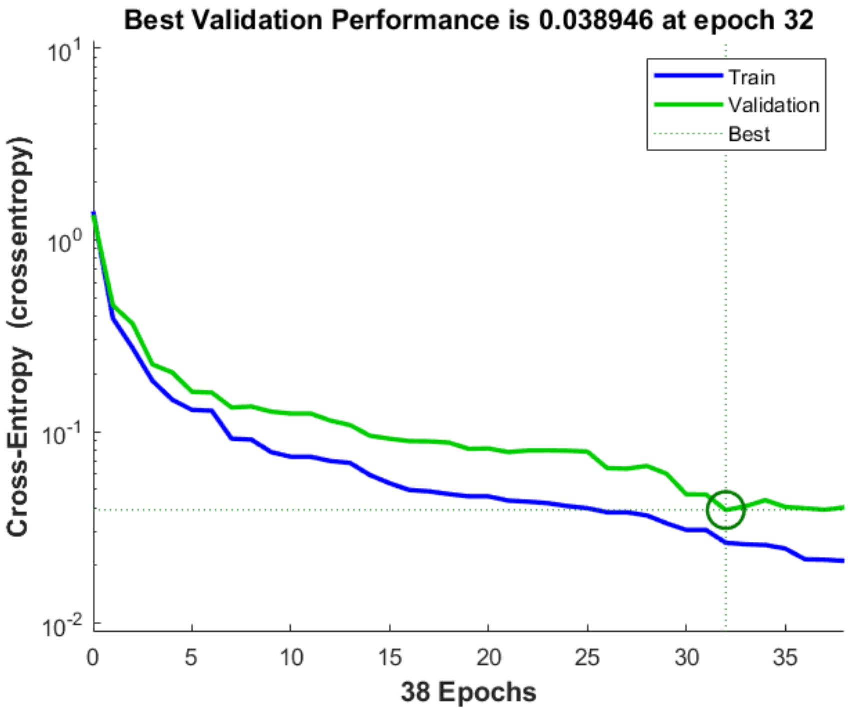
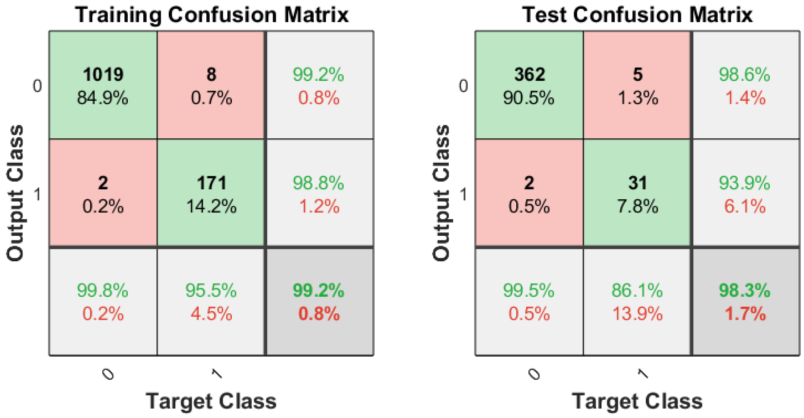


Figure 42: Trend of the neural net training

After the training, the neural network is verified using test set data. The following figure shows the confusion matrices of both training data and test data:



Training Confusion Matrix

Output Class \ Target Class	0	1	
0	1019 84.9%	8 0.7%	99.2% 0.8%
1	2 0.2%	171 14.2%	98.8% 1.2%
	99.8% 0.2%	95.5% 4.5%	99.2% 0.8%

Test Confusion Matrix

Output Class \ Target Class	0	1	
0	362 90.5%	5 1.3%	98.6% 1.4%
1	2 0.5%	31 7.8%	93.9% 6.1%
	99.5% 0.5%	86.1% 13.9%	98.3% 1.7%

Figure 43: Training and test confusion matrix

On the confusion matrix plot, the rows correspond to the predicted class (output class) and the columns correspond to the true class (target class). The diagonal cells correspond to observations that are correctly classified while the off-diagonal cells correspond to incorrectly classified observations. Both the number of observations and the percentage of the total number of observations are shown in each cell.

Training performance are obviously better of the test one, but the high precision of the test proves that there are no overfitting concerns. The result can be considered satisfactory, other attempts have been performed changing the hidden layers size but they have not lead to significant improvements.
Another program has been chosen in order to verify the accuracy of the neural network classification. The new dataset (PLS, group 07, program 011) has 2454 total points, 2156 without splash and 298 with. The 80% of these points are used as training and validation set, the remaining 20% as test set. A neural network with a 280 input layers (in this program, the welding process lasts 280 *ms*), two hidden layers of 100 and 20 and the single output layer. With the same previous procedure, the training and the test are carried out and these are the resulting confusion plot:

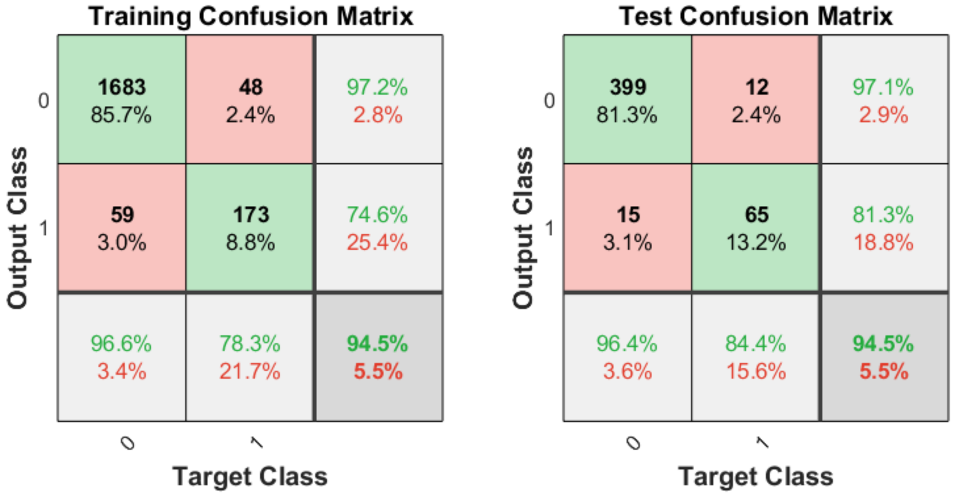


Figure 44: Confusion plot of splash classification

The accuracy is acceptable but not excellent. The number of false negative (those points that are evaluated as safe while the target is on warning), that is the worst case, and of false positive (evaluated as warning while being safe) is too high with respect to the true positive (correctly evaluated as splash points). This means that the classification accuracy can be affected by the program of the welding guns. It is possible to try another way to both improve the classification and lighten the computational load: the introduction of statistical descriptors.

4.1.1 Statistical descriptors

Statistical descriptors are indices that can synthetically and effectively describe a set of data. They belong to the field of the descriptive statistics and can be used both with continue and discrete variables.

They can be divided into:

- position indices:
    - mode,
    - median,
    - mean.
- dispersion indices:
    - standard deviation,
    - variance.
- shape indices:
    - skewness,
    - kurtosis.

Position indices (also known as central trend measures) identify, in different ways, the central element of the distribution.
Dispersion indices evaluate how much data deviates from the central element of the distribution.
Finally, shape indices consider the shape of the distribution with respect to a normal (or Gaussian) distribution. In particular, skewness indicates how much the distribution is asymmetric and kurtosis how it is flat. Mathematically, the skewness is defined as:

$$skewness = \frac{m_3}{m_3^{3/2}} \tag{26}$$

while kurtosis:

$$kurtosis = \frac{m_4}{m_2^2} \tag{27}$$

where m_k is the central moment of order k:

$$m_k = \sum_{i=1}^{n} (x_i - \mu)^k \tag{28}$$

x_i are the values of the dataset and μ is the mean value.

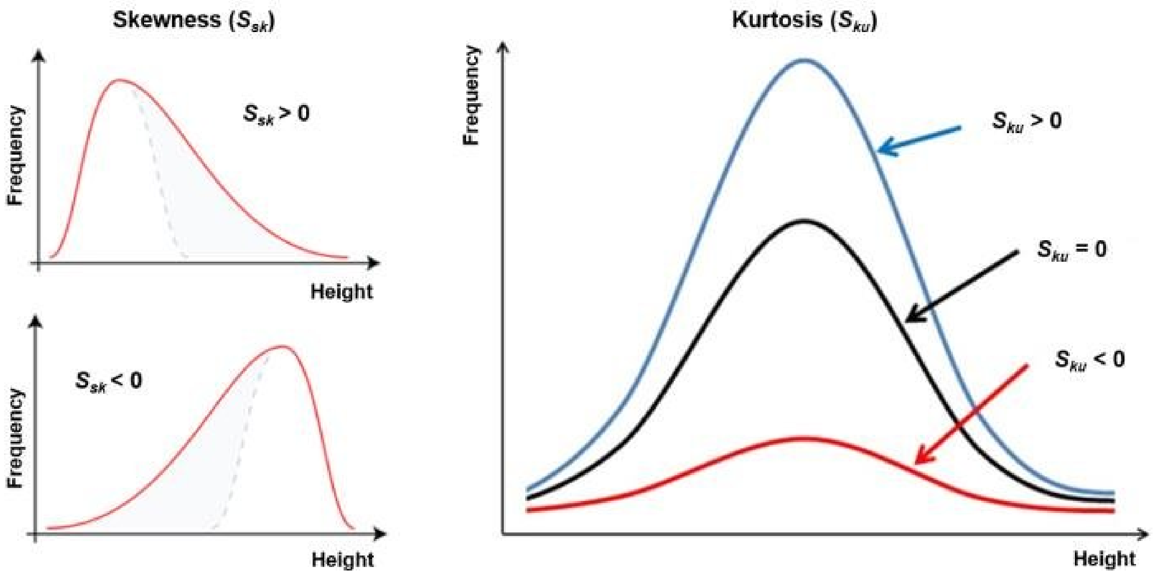
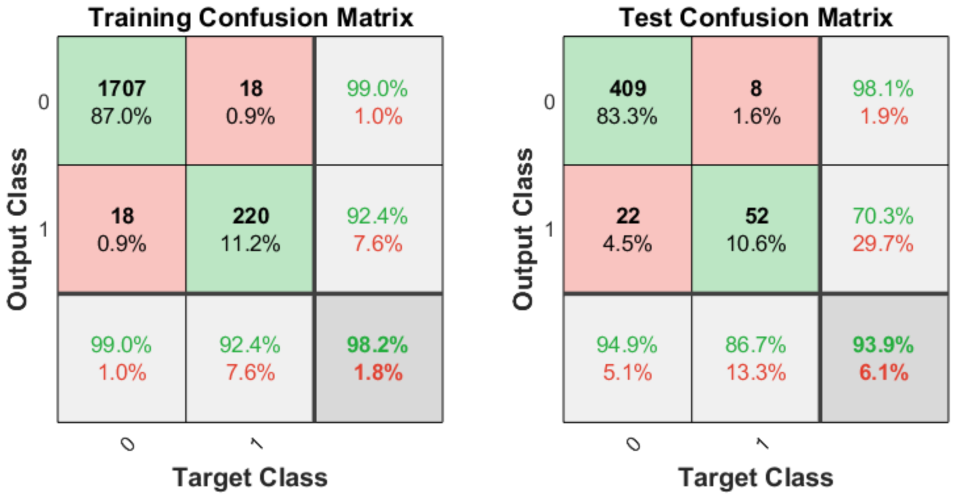


Figure 45: Skewness and kurtosis graphically explained

So, these indices are able to describe the distribution and can be used instead of taking in consideration all the measured values. A new neural network has been implemented with a new set of inputs. These inputs include:

- statistical descriptors: mean, mode, median, standard deviation, $3rd$ order and $4th$ order central moment, skewness, kurtosis, max and min value of a welding point.
- other indices automatically computed from the company software: initial and final resistance values, instants of max and min values, rise and descent time, growth and degrowth rate.

A total of 18 inputs for each welding point is thus selected. The hidden layer has been set of 20 layers and the output is the usual one, 0 for safe condition and 1 for splash detection. The computational load is much lighter than the previous one and the result is better:



Training Confusion Matrix

Output Class \ Target Class	0	1	
0	**1707** 87.0%	**18** 0.9%	99.0% 1.0%
1	**18** 0.9%	**220** 11.2%	92.4% 7.6%
	99.0% 1.0%	92.4% 7.6%	**98.2%** **1.8%**

Test Confusion Matrix

Output Class \ Target Class	0	1	
0	**409** 83.3%	**8** 1.6%	98.1% 1.9%
1	**22** 4.5%	**52** 10.6%	70.3% 29.7%
	94.9% 5.1%	86.7% 13.3%	**93.9%** **6.1%**

Figure 46: Confusion plot of splash classification with statistical descriptors

In particular, the number of false negatives is smaller than the previous one, this means that the algorithm with these input recognize in a better way the presence of the splash. The number of false positive is higher. This can be due to a WMS issue: this software detects the presence of splash on the percentage drop (20%) in resistance in two consecutive samples. Its sample time is 4 ms, and sometimes it is possible that despite the splash shows up, the software does not detect this percentage drop and classifies the welding point as not affected by splash. For example, the following figure displays two points that, according to WMS, are not affected by splash:

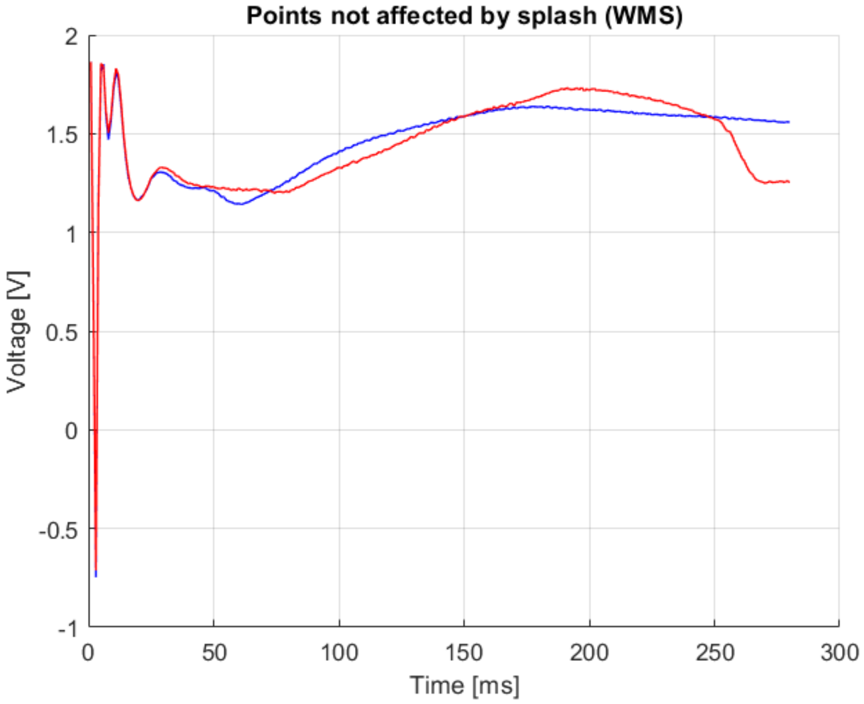


Figure 47: No-splash points, according to WMS

Actually, the red voltage trend shows the characteristic decrease of the splash, and the neural network classifies it as a warning point, falling in a false positive classification.
In conclusion, it can be affirmed that the neural network classification is able to recognize the presence of the splash. The algorithm accuracy can be improved by increasing the dimension of the dataset, so that more splash examples can feed the network improving its learning ability. Another important step can be the measurement of new critical variables that may deal with the splash disturbance (e.g., pressure force, temperature, incidence angle between electrodes and sheets, etc.) and their consequent insertion in neural network inputs.

5 Splash prediction

5.1 Splash prediction with classification

Since the neural network classification approach gave good results, it is possible to try a prediction analysis with this methodology. An important consideration is that a splash is unlikely to occur in the initial phase of the welding process because the melting starts from the point of contact of the sheets, where there is greater resistance, while the splash occurs when the weld core expands until it comes into contact with the electrodes. The idea is that the various causes that lead to the splash can in some ways affect the electrical variables before its physical manifestation and that machine learning algorithms can find out numerical pattern to detect these deviations. With this hypothesis, a first try of prediction is made considering only the transient (about 40 ms) of the current, voltage and resistance values (recalling that the last is calculated and not measured) in order to find out if it contains sufficient information to predict an incoming splash.
The first dataset (PLD, group 02, program 005) is again chosen, since the previous classification leads to good results. Statistical descriptors are thus computed for the variables, taking into account only the blanking time. Ten statistical descriptors for each of three variables, the eight WMS indices cannot be used because they are computed at the end of the welding process and contains information on the whole trend. The input layers is a 30 neurons layer, while the output layers is the usual binary one. As regard the hidden layers, different sizes have been trained and the best performance is obtained with 50 layers one:

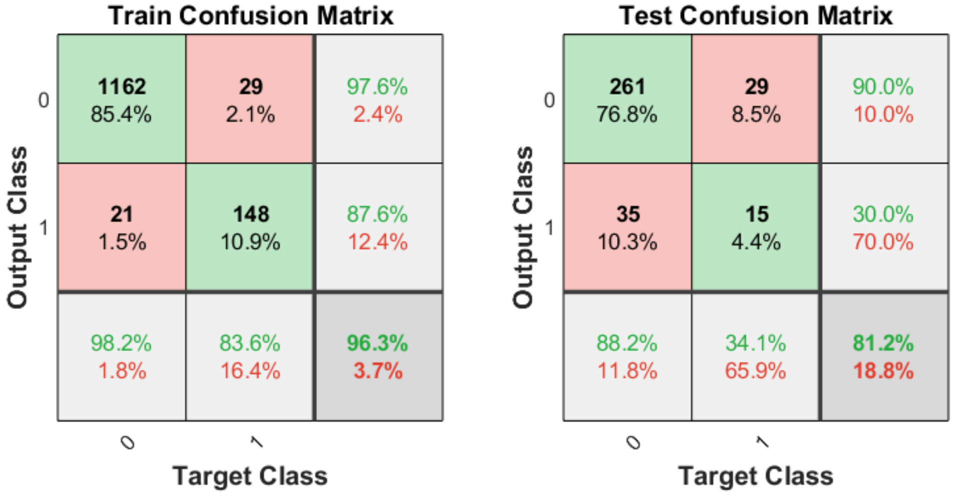


Figure 48: Confusion plot of splash prediction with classification

The train confusion matrix gave good results, but the test one is not acceptable. This means that the overfitting problem occurred: the algorithm adapts too much to the training set and loses the capability to evaluate new examples. In order to avoid this

situation, it is possible to decrease the number of features or the dimension of the hidden layers size. Both of these solutions have been implemented, but unfortunately they did not lead to a better result.
At this point the other dataset has been chosen (PLS, group 07, program 011) and with the same settings, a new algorithm has been trained and tested:

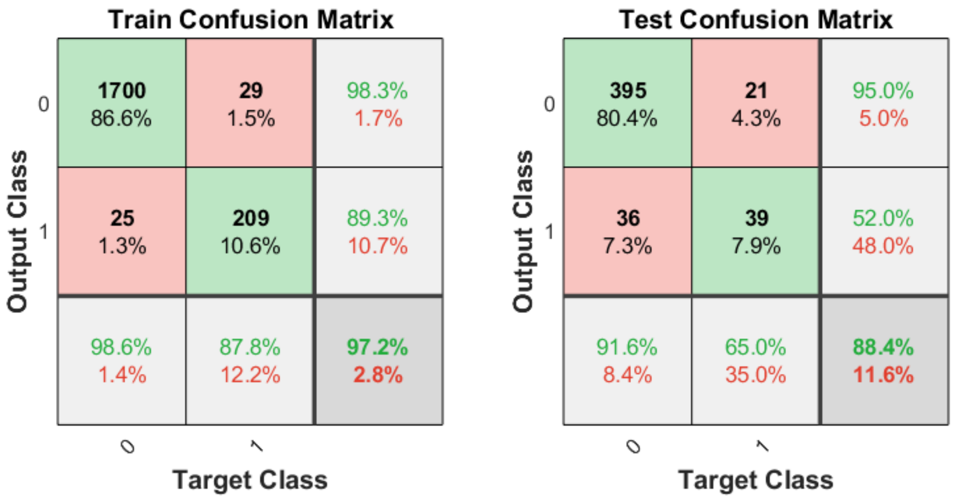


Figure 49: Confusion plot of splash prediction with classification

Also in this case the training performance is significantly better than the test performance, although it is possible to point out a slight improvement with respect to the previous one. In conclusion, the algorithm did not find any global common pattern or difference among the data that precede or not the splash. On the other hand, there is a certain percentage of correctly predicted splash that stresses a possible relationship between the structural causes that lead to splash and the electrical variables.

5.2 Splash prediction with time-series algorithms

In addition to the classification, a neural network can be designed to obtain as output a sequence of values. Moreover, some algorithms that can be used to model and estimate a system can be modified to obtain an algorithm that predict the values of the variables of a time-series. The most known algorithms to be effective with time series are the NARX neural network, which has already been widely discussed, and the Long-Short Term Memory algorithm.

5.2.1 LSTM algorithm

In order to understand the LSTM algorithm, a little introduction of recurrent neural network (RNN) is needed. A recurrent neural network is a deep learning network structure that exploits the information of past values to improve the performance of the network on current and future inputs. RNNs contains a hidden state and loops, in which the looping structure allows the network to store past information in the hidden state and operate on sequences. In fact, this algorithm is mainly used with sequential data of varying length such as natural language processing, signal classification, video analysis and, in general, time-series.

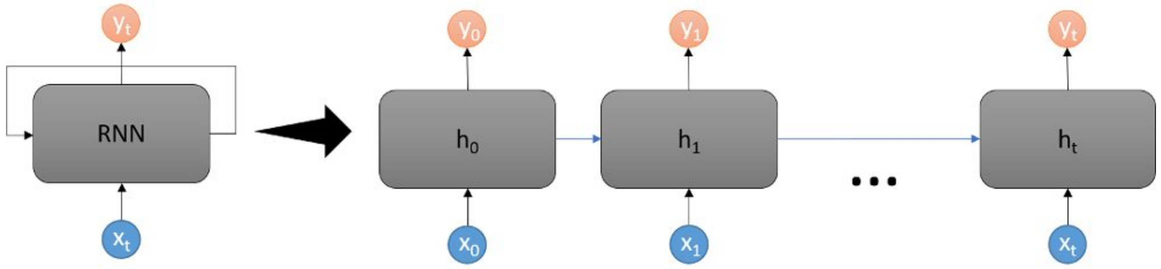


Figure 50: Recurrent neural network

This network is able to learn short-term dependencies because it has two sets of weights, one for the hidden state vector and one for the inputs. During training, the network learns weights for both the inputs and the hidden state, and when implemented, the output is computed on the current input and on the hidden state, which is based on previous inputs. This type of network has a limit, the vanishing gradients often leads to capture short-term dependencies while long-term ones are not evaluated. On the other end also exploding gradients may occur, causing the error to grow drastically with each time step.

Long short-term memory networks have been designed to solve these issues by using the gates to selectively retain information that is relevant and forget information that is not relevant. This algorithm, in fact, uses additional gates to control what information in the hidden cell makes it to the output and the next hidden state. Lower sensitivity to the time gap makes long short-term memory networks better for analysis of sequential data than simple RNNs.[12]

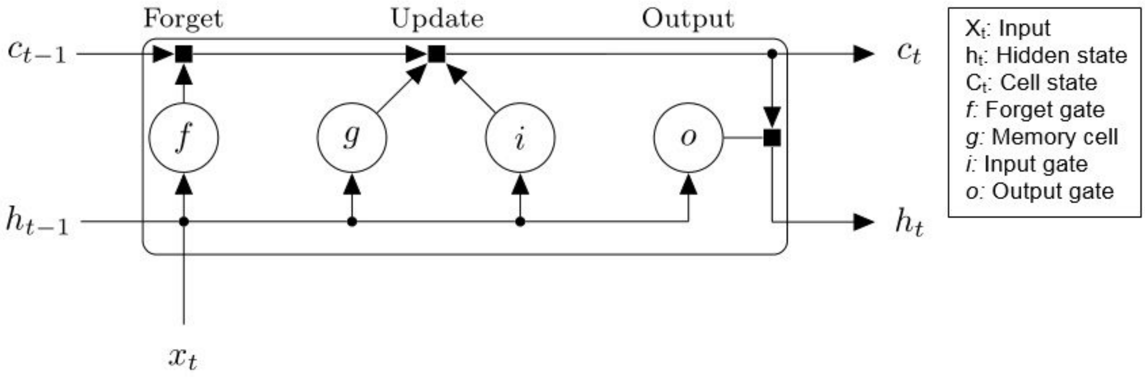


Figure 51: Long short-term memory network

5.2.2 Forecasting with LSTM algorithm

A long short-term memory algorithm has been designed in order to be trained with real welding points (PLD, 002, 05). In this network, the input is the voltage at time *t*, the output is the voltage at time *t+1*, the LSTM hidden layer has a size of 200. So, during the training, the network computes the output instant by instant while the hidden layer tries to find long and short term relationship between the fed voltage values. Since the voltage is measured during the welding process, when the test is performed it is possible to give the real output values as network input to predict the subsequent step.

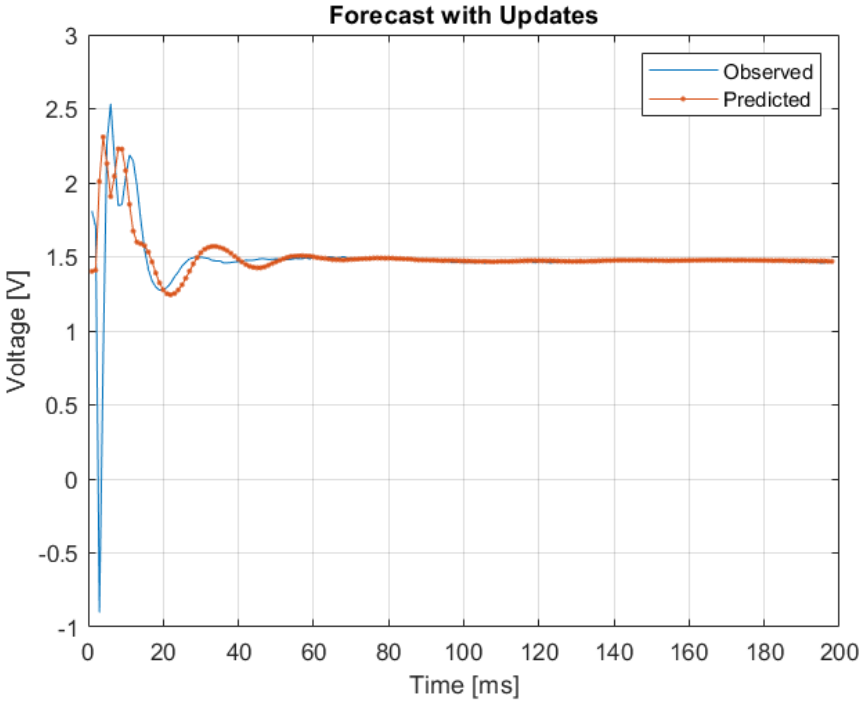


Figure 52: Voltage prediction through LSTM algorithm, point without splash

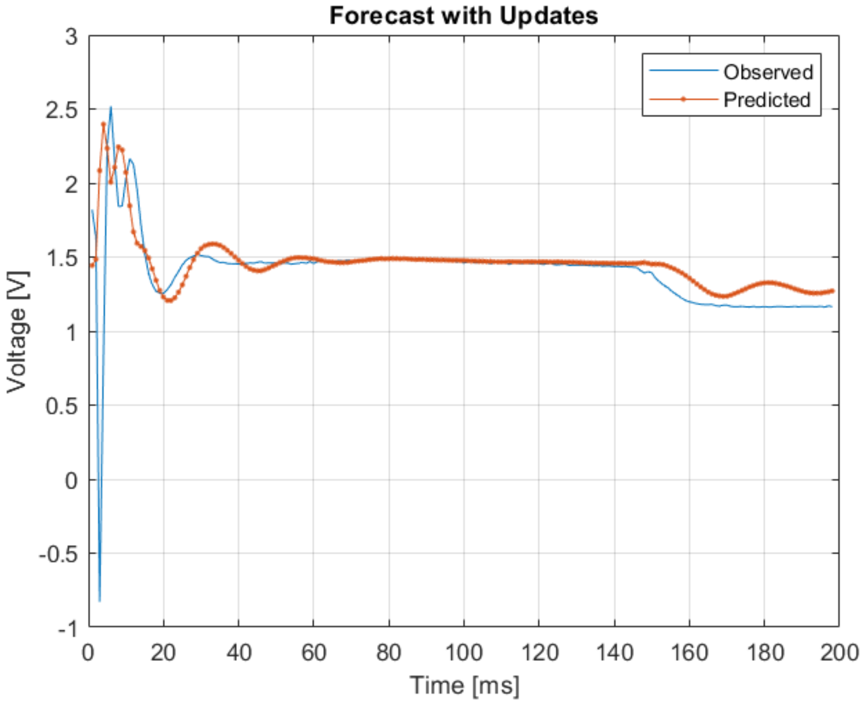


Figure 53: Voltage prediction through LSTM algorithm, point with splash

The prediction is not acceptable, because after the blanking time the output of the network is just able to follow the real output with the values that are fed it back, and not to predict. Moreover, when the splash shows up the algorithm is not even able to recover the real position.

Since the blanking time is not interesting from the splash point of view, it is possible to select the significant part of the data removing the first 40 ms from each welding spot. A new training is performed, and the obtained results are clearly better:

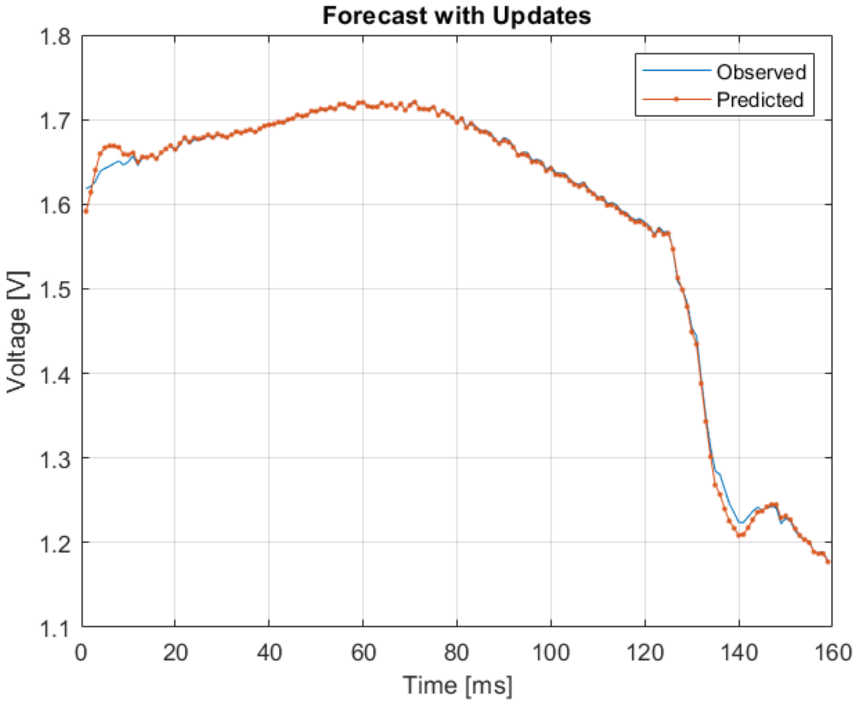


Figure 54: Voltage prediction through LSTM algorithm without blanking

The blanking acquisitions introduce a wider range of values and more oscillations, neglecting them the network performs better. Anyway, a 1 *ms* prediction is not useful for predictive maintenance purposes, it is a small time window to detect splash evidences and especially to act in advance to avoid it. Recalling that the WMS has a 4 *ms* sampling time with which it reacts to splash, it can be a reasonable idea to forecast with this prediction horizon. The input is still the voltage, the output is the voltage four instant later.

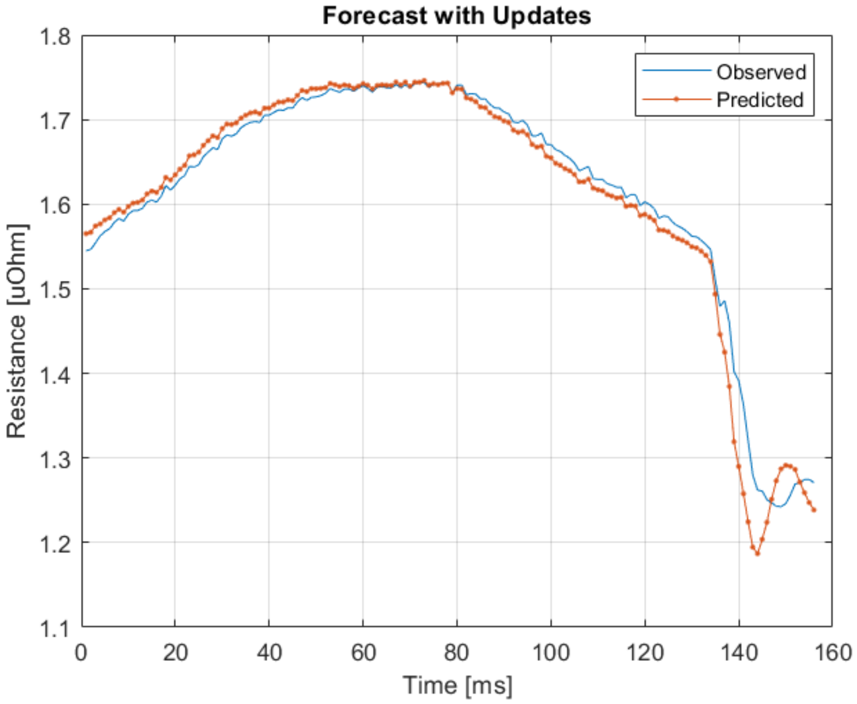


Figure 55: Voltage observations and 4 ms prediction, without blanking

The orange line should precede the blue one by 4 instants, but it does not happen. Anyway, there is a short forecast range, as it is highlighted in this image:

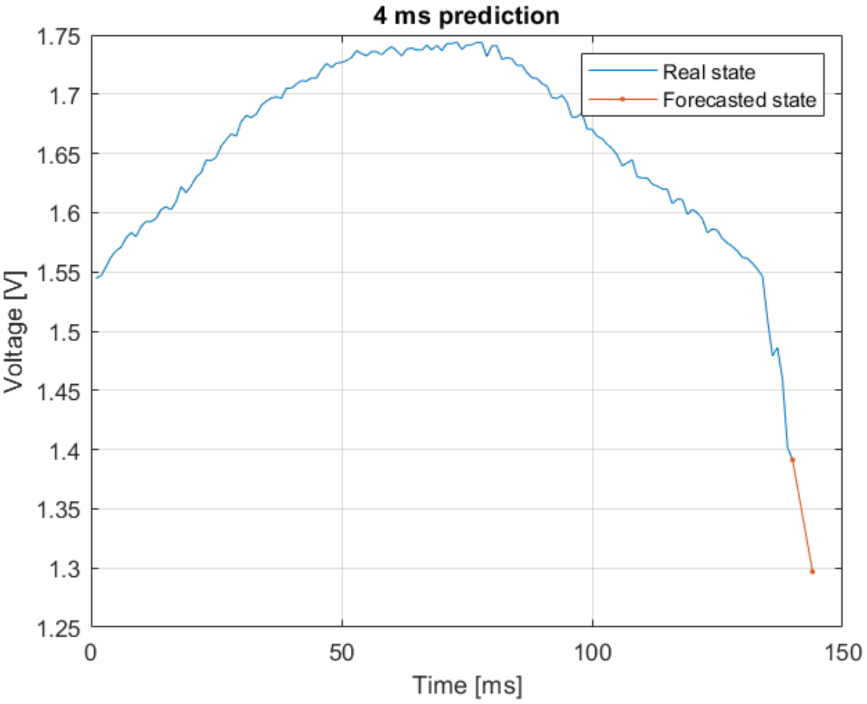


Figure 56: Real and forecasted state of the voltage

Simulating a real situation, at instant 180 (the last blue line value, recalling that the blanking time has been removed, corresponds to 140 in the figure) there is a probable splash in progress, the network output (that is 4 ms forward) shows that the voltage curve will continue to decrease.
A last attempt is performed increasing again the prediction horizon to 8 ms:

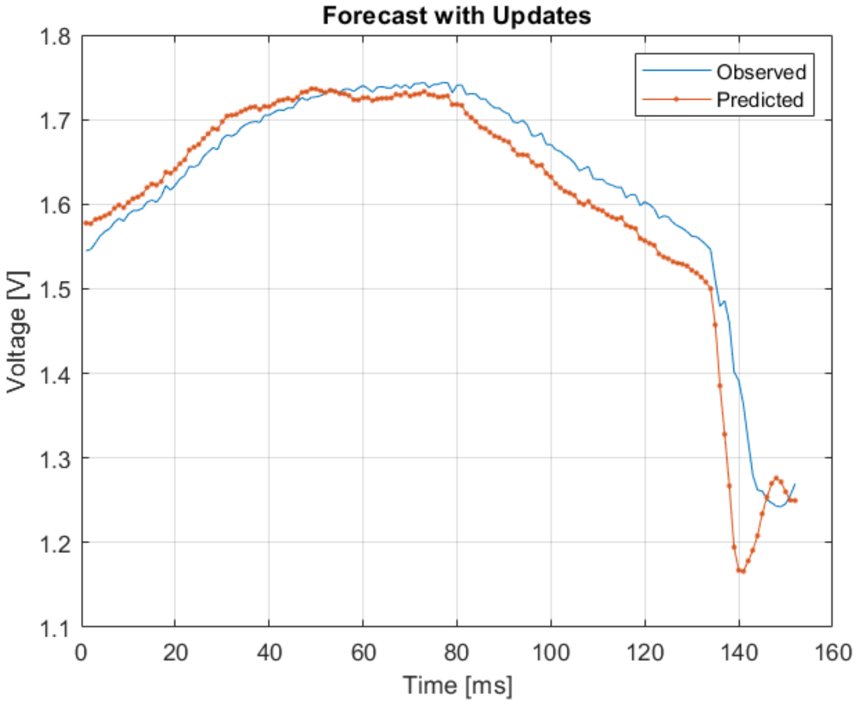


Figure 57: Voltage observations and 8 ms prediction, without blanking

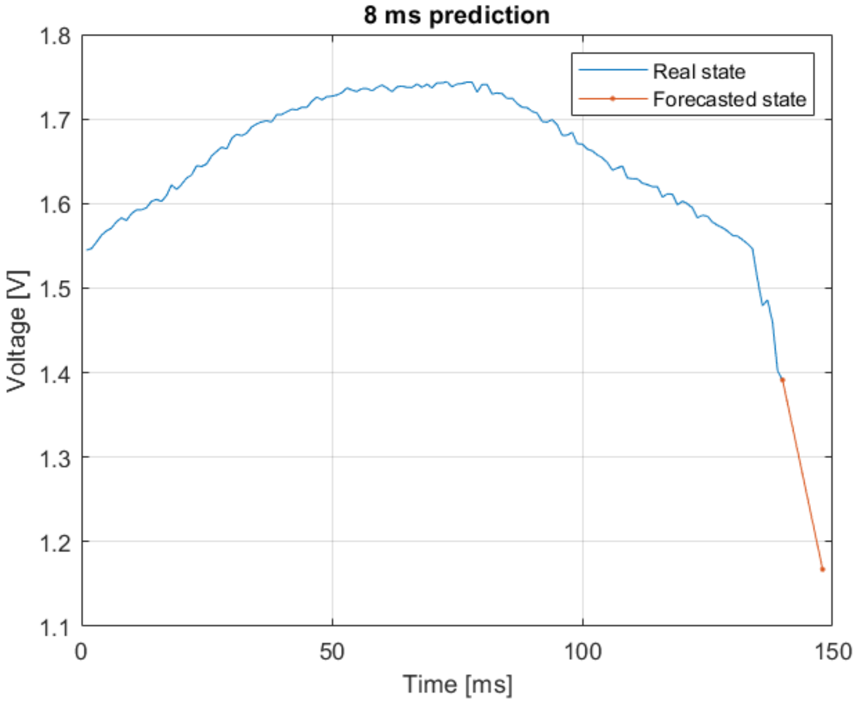


Figure 58: Real and forecasted state of the voltage

Again, in the first figure it is possible to see that there is a forecasting range that not correspond to the prediction horizon. In conclusion, as it is possible to see in the second figure, this algorithm with such information is not able to predict an incoming splash but, once it has started, on the basis of the predicted output it could be possible to warn the WMS faster than the current strategy does.

5.2.3 Forecasting with NARX algorithm

NARX algorithm has been used in order to find a model of the welding gun system. With some changes to network settings, the goal of using it for prediction is achievable. The network has to be trained approximately in the same way as before: selection of the dataset, of the delayed states (recalling that the past values of the input and the output are given as network input) and in addition the prediction horizon. A network with 5 delayed states and a 5 *ms* prediction horizon has been designed and trained. This means that, when the network is used to predict, the first network output comes out after the first five measured values of the variables and represents the prediction five instants forward. Basically, the first network output represents the voltage at the 10*th ms* of the welding process. After the training, a test is performed using other points of the same program:

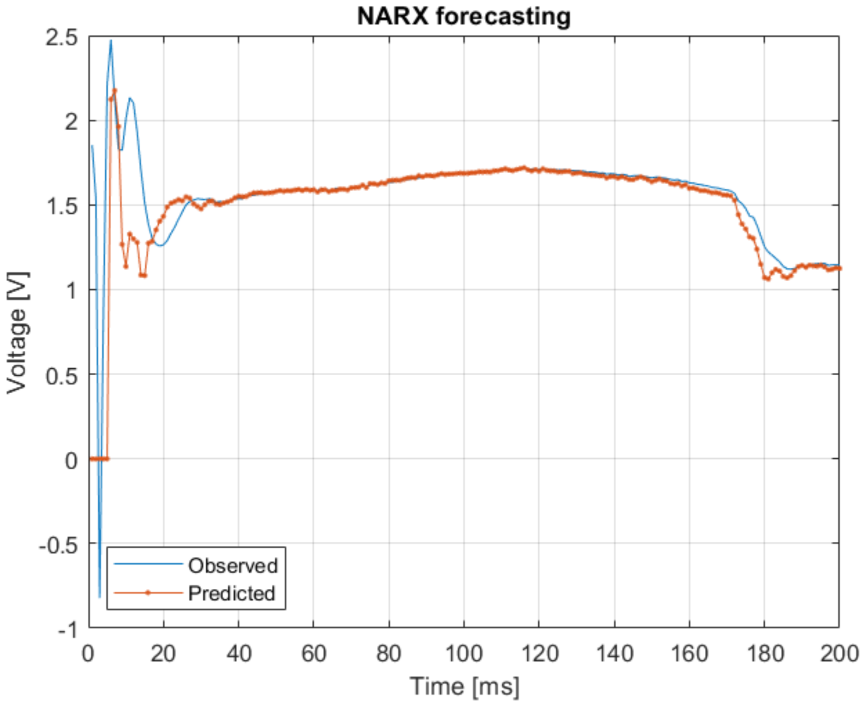


Figure 59: Voltage prediction through NARX algorithm, point with splash

The result is similar to the LSTM one. After the first 5 *ms*, the orange line should precede the blue one of other 5 *ms* but this does not happen. However, as soon as the splash occurs, the predicted output has a little range of prediction that can result in a faster

response of the control software.
Also in this case, neglecting the blanking time could lead to achieve better results. Removing these values from data and re-training the network, the test is displayed in the following figure:

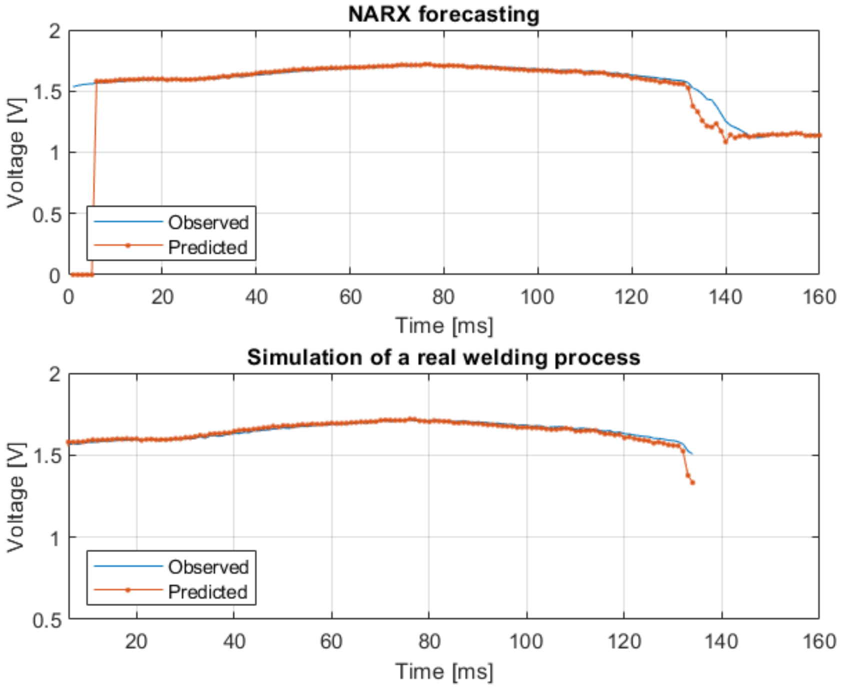


Figure 60: Simulation of a real welding process, 5 ms prediction

In the top subplot it is evident that the voltage descent has a larger prediction range. In the bottom subplot a real situation has been simulated: the information about the variables fed to the network does not have clear evidence about a splash, but the predicted output shows a decreasing slope typical of the splash.
A last attempt is performed widening the prediction horizon to 8 *ms*. Another test point is selected and fed to the new trained network:

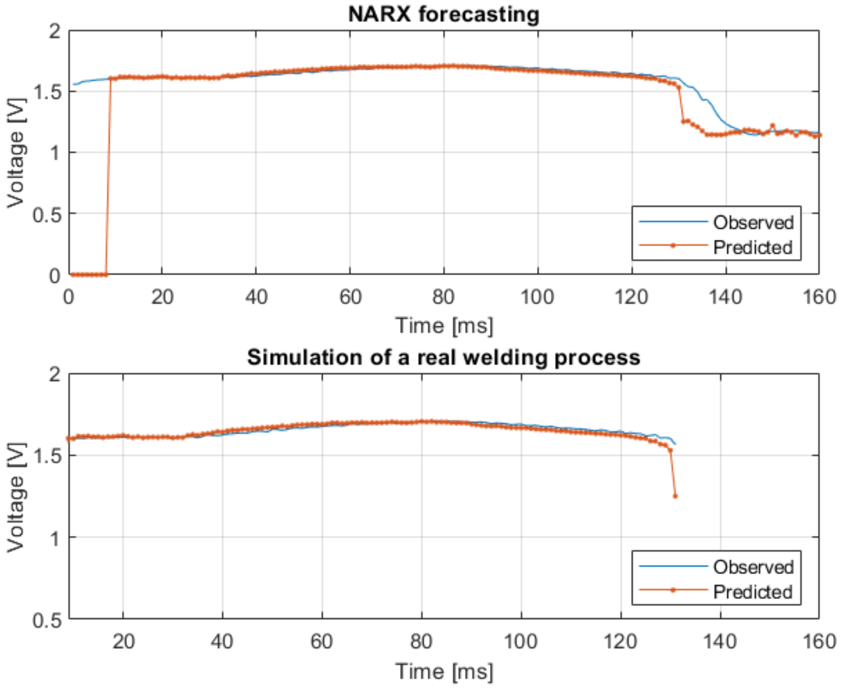


Figure 61: Simulation of a real welding process, 8 ms prediction

Even with these settings it seems that the network is not very able to predict the splash with a large advance. However, the reaction to a minimal measured decrease of slope lead to a great prediction of the slope, as can be seen in the second subplot.
In conclusion, these algorithms can be implemented to react in a faster way to a splash. The measured electrical variables do not contain enough information to predict an incoming splash, but they can be exploited to limit the damages.

6 Conclusion

The industry 4.0 arises from the ongoing industrial automation that in the last years has exploited the new available technologies to improve work conditions, to create new business model and to increase productivity and quality.

Large scale machine-to-machine communication and the internet of things (IoT) are integrated to make the machine smarter and smarter, able to self-monitor and to analyze issues without the need for human intervention.

For this reason, it is not a long shot to say that predictive maintenance will become, in the next years, a widely used technology in every industrial context.

This thesis work represents the first steps towards the development of a predictive maintenance system for welding guns. Above all, the ISI-Welding Systems main purpose was to find out the feasibility of this approach in order to provide more suitable equipment to pursue this strategy in the near future. Unfortunately, the adverse pandemic period has made a direct contact between the graduate students and the company difficult, and it was not possible to implement and test the developed algorithms on a real welding gun.

To sum up, good results have been obtained under different aspects:

- **Model**: the studies carried out for the development of a suitable model provided a deeper knowledge of the welding process and, in particular, useful clues for the variables to focus on for future measurements. On the basis of current and voltage, the estimated physical model gives a good approximation of the real phenomenon, it will be easier to improve the performance with the introduction of new variables and with the widening of the database.

- **Predictive algorithms**: in general, encouraging results have been obtained with machine learning algorithms, proving the feasibility of the predictive maintenance approach.

  As regard splash classification, the introduction of statistical descriptors represent a great advantage in terms of accuracy performance and computational load. In particular, they are useful when the welding process is over and the whole trend of the variables is available, so this strategy is not very advantageous in term of prediction but can be exploited for an intelligent data-driven splash classification.

  As regard splash prediction, algorithms that work with real-time measurement (LSTM and NARX) have proven to be functional for a fast detection of the splash. Their implementation on real welding guns can show if they can allow an improved reaction to this disturbance rather than the present strategy, allowing to consistently avoid bad quality resistance welding spots and all the inconveniences that can arise from an enduring splash.

6.1 Future developments

The future developments regard the implementation of these algorithms on real welding guns and the fulfillment of the conditions useful to enhance and facilitate the studies towards the predictive maintenance strategy. As time goes on, even more collected data will enlarge the available databases, that is a fundamental condition to improve data-driven algorithms. Moreover, the introduction of new crucial variables, such as actuation force of the electrodes, welding temperature or heat and visual detection of the the positioning of the electrode axes in relation to the sheets can significantly improve the algorithm performances and lead to the development of a definitive welding gun model. It would be interesting a crossed study between the splash disturbance and the electrode dressing, that are the issues analyzed in a deeper way in this and the other team component works. The predictive maintenance is definitely an advantageous strategy that can differentiate and increase the status of a company with respect to the other ones. The efforts towards the achievement of this goal imply changes in every aspect of the working conditions, from the design of machines for an easier supervision to the additional resources for data acquisition and analysis, but they will result in a greater working environment in terms of human operator safety, resource management and quality of the product.

References

[1] *Story of the ISI-Welding Company*, http://en.isi-gf.com/about.aspx?TypeId=1&FId=t1:1:1

[2] Production engineering, https://www.resistanceweldsupplies.com/weld-help/cause/inadequate-or-no-gun-equalization.html#:~:text=An%20equalizing%20system%20compensates%20for,2

[3] Erik Baldoin, *Evoluzione delle pinze robot per saldatura a punti*, Politecnico di Torino, 2009

[4] Luciano Furlanetto, *Manuale di manutenzione degli impianti industriali*, FrancoAngeli, 2003

[5] Nakajima, 1984

[6] L. Fedele, L. Furlanetto, D. Saccardi, *Progettare e gestire la manutenzione*, McGraw Hill, 2004

[7] G. De Nicolao, R. Scattolini, *Identificazione Parametrica*, Dipartimento di informatica e sistemistica dell'Università degli Studi di Pavia

[8] Luigi Fortuna, Salvatore Graziani, Alessandro Rizzo, Maria Gabriella Xibilia, *Soft sensors for monitoring and control of industrial processes*, Springer Science & Business Media

[9] Matlab, *Time Series NARX Feedback Neural Networks*, https://it.mathworks.com/help/deeplearning/ug/design-time-series-narx-feedback-neural-networks.html

[10] Ruiji Sun1, Wangling Yu, Haiyan H. Zhang, Manohar Das, Qingyou Han, *Mathematical Modeling of Resistance Spot Welding*, Journal of Multidisciplinary Engineering Science and Technology (JMEST)

[11] R. Adami, *La saldatura a resistenza*, Università degli Studi di Padova

[12] Matlab, *Recurrent neural network*, https://it.mathworks.com/discovery/rnn.html