H2020 – ICT-13-2018-2019

MUSKETEER

Machine Learning to Augment Shared Knowledge in Federated Privacy-Preserving Scenarios (MUSKETEER)

Grant No 824988

D2.1 Industrial and Technical Requirements

March 19

Imprint

Contractual Date of Delivery to the EC: **31 March 2019**

Author(s): **Joao Correia (B3D)**
Participant(s): **Harry Hatzakis (B3D), Chiara Napione and Lucrezia Morabito (COMAU), Giacomo Fecondo (FCA), Christina Kotsiopoulou and Petros Papachristou (HYGEIA), Susanna Bonura (ENG)**
Reviewer(s): **Giacomo Fecondo (FCA), Antoine Garnier (IDSA), Maria-Irina Nicolae, Karen McKenna, Gal Weiss (IBM)**

Project: **Machine learning to augment shared knowledge in federated privacy-preserving scenarios (MUSKETEER)**

Work package: **WP2**
Dissemination level: Public
Version: **1.1**

Contact: **Joao Correia – jcorreia@biotronics3d.com**
Website: **www.MUSKETEER.eu**

Legal disclaimer

The project Machine Learning to Augment Shared Knowledge in Federated Privacy-Preserving Scenarios (MUSKETEER) has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824988. The sole responsibility for the content of this publication lies with the authors.

Copyright

Executive Summary

The deliverable D2.1 – Industrial and Technical Requirements is the first result the task T2.1 Industrial scenarios objectives refinement and technical requirements. It envisages to elicit the end user requirements derived from the needs of the two industrial scenarios considered within the project. These scenarios will provide datasets that have been deeply analysed in order to identify all the technical challenges implied with the processing of such datasets. The deliverable presents a complete specification of the functional and non-functional user's requirements, technical requirements and privacy operation models that should compose the MUSKETEER data platform.

Document History

Version	Date	Status	Author	Comment
0.1	15 Feb 2019	Initial structure	Joao Correia	First draft
0.2	28 Feb 2019	Medical Use Case	B3D	
0.3	04 Mar 2019	Manufacturing Use Case	COMAU/ FCA	Update
0.4	08 Mar 2019	Medical Use Case	HYGEIA	Update
0.5	11 Mar 2019	Conclusions	B3D	Update
0.6	12 Mar 2019	For internal review	B3D	Update
0.7	14 Mar 2019	Review improvements	All	Update
0.8	18 Mar 2019	Reviewers inputs	Giacomo Fecondo (FCA), Antoine Garnier (IDSA)	Update
0.9	20 Mar 2019	Inputs for: methodology, Technical Requirements, conclusions	Susanna Bonura (ENG)	Update
1.0	27 Mar 2019	Final Version	All	
1.1	28 Mar 2019	Finalization	Gal Weiss	Update tables

Table of Contents

LIST OF FIGURES 4
LIST OF TABLES 4
LIST OF ACRONYMS AND ABBREVIATIONS 5
1 INTRODUCTION 6
1.1 Purpose 6
1.2 Related Documents 6
1.3 Document Structure 6
2 REQUIREMENTS ELICITATION METHODOLOGY 7
2.1 Description of the methodology 7
2.2 Requirements Definition 8
3 CHARACTERISATION OF THE INDUSTRIAL SCENARIOS 9
3.1 Smart Manufacturing 9
3.1.1 User Stories 14
3.1.2 Data Characterisation 16
3.1.3 The Data Flow 19
3.1.4 Privacy Operation Modes 19
3.2 Health – Medical Imaging 20
3.2.1 User Stories 23
3.2.2 Data Characterisation 24
3.2.3 The Data Flow 28
3.2.4 Privacy Operation Modes 29
3.3 MUSKETEER User Goals 31
4 SPECIFICATION OF REQUIREMENTS 31
4.1 User Roles 32
4.2 Industrial Users' Roles in MUSKETEER Platform 32
4.3 MUSKETEER user requirements 33
4.3.1 Functional Requirements 33
4.3.2 Non-functional requirements 38
4.4 MUSKETEER Technical Requirements 39
5 CONCLUSION 43
6 REFERENCES 43

List of Figures

Figure 1. MUSKETEER project methodology 7
Figure 2. Requirements elicitation in MUSKETEER 8
Figure 3. Welding Robot 10
Figure 4. Welding process logical representation 11
Figure 5. Welding gun image 13
Figure 6. Spot-welding time cycle 14
Figure 7. Smart Manufacturing local data collection 18
Figure 8. Smart Manufacturing architecture 19
Figure 9. POM PORTHOS 20
Figure 10. Prostate image visualisation and report 22
Figure 11. PI-RADS assessment (from radiologyassistant.nl) 26
Figure 12. PI-RADS segmentation (from PI-RADS™ v2) 26
Figure 13. Medical Imaging data flow 29
Figure 14. POM RICHELIEU 30

List of Tables

Table 1 Smart manufacturing user stories 14
Table 2 Welding data characterization 16
Table 3 Health – medical imaging user stories 23
Table 4 Available data 27
Table 5 Average frequency of data 27
Table 6 MUSKETEER User roles 32
Table 7 Mapping between industrial users' roles and MUSKETEER user roles 33
Table 8 Functional requirements 33
Table 9 Non-functional requirements 38
Table 10 Technical requirements 39

List of Acronyms and Abbreviations

Abbreviation	Definition
DICOM	Digital Imaging and Communications in Medicine
HL7	Health Level Seven International
PI-RADS	Prostate Imaging – Reporting and Data System
kA	Kiloampere
daN	Dekanewton
s	second
ms	milliseconds
RSW	Resistance spot welding
POM	Privacy Operation Mode
J	Joule

1 Introduction

1.1 Purpose

This document presents the end user and technical requirements for MUSKETEER platform derived from the needs of the two industrial scenarios considered within the project. The datasets provided by these scenarios, from distinct areas of smart manufacturing and healthcare, have been analysed in order to identify all the technical challenges implied with the processing of such datasets.

1.2 Related Documents

The deliverable D2.1 – Industrial and Technical Requirements is the first result of WP2 and in particular of the task T2.1 Industrial scenarios objectives refinement and technical requirements. It constitutes a basis for the remaining work packages regarding architecture, development, testing and validation of the MUSKETEER platform. More in detail, the document presents a complete specification of the user stories in Smart Manufacturing and Health scenarios, business requirements coming from the two scenarios (grouped in user functional and non-functional requirements), and technical requirements to meet such business requirements and drive technical developments in WPs 3, 4, 5 and 6, and in WP7 on use case piloting and validation.

In addition, the deliverable D2.1 will be considered as a main input for the description of the technical and domain business-specific KPIs that will be used for validating the MUSKETEER data platform (deliverable D2.3).

1.3 Document Structure

The structure of the document is as follows:

In Section 2, the requirements identification and elicitation methodology are defined. At first, an overview of the adopted development processes, instruments, roles and methods is provided. Moreover, in this section the User Stories definition process is presented, as well as the requirements definition process.

In Section 3, for each scenario, a general description of the use cases, user stories, current operation, existing data, origins and flows of data and candidate POMs are described in detail.

In Section 4, the detailed specifications of user and technical requirements for the MUSKETEER platform are presented based on the deep analysis of the industrial scenarios. First, the roles

of users in MUSKETEER platform are described and mapped to the industrial users identified in the industrial scenarios. User requirements are specified based on the user stories description and categorised into functional and non-functional requirements. Finally, the technical requirements translate the user requirements into implementation-oriented requirements.

Section 5 concludes the deliverable. It outlines the main findings of the deliverable which will guide the future research and technological efforts of the consortium.

2 Requirements Elicitation Methodology

2.1 Description of the methodology

According to the MUSKETEER Description of Work, the project consists of five research stages (Figure 1). Stage 1 focuses on the identification of requirements in different scenarios (WP2). Stage 2 concerns the design of the different modules which form the MUSKETEER platform (WP3, WP4, WP5 and WP6). Stage 3 will convert previous designs in real software developments. In Stage 4 the behaviour of the platform from different perspectives (scalability, computational efficiency, security and data value contribution) is assessed. Finally, in Stage 5 the MUSKETEER platform we be validated through the use cases. The different stages are overlapped in time, running in parallel and iteratively on an agile manner to ensure feedback, coherency and consolidation.

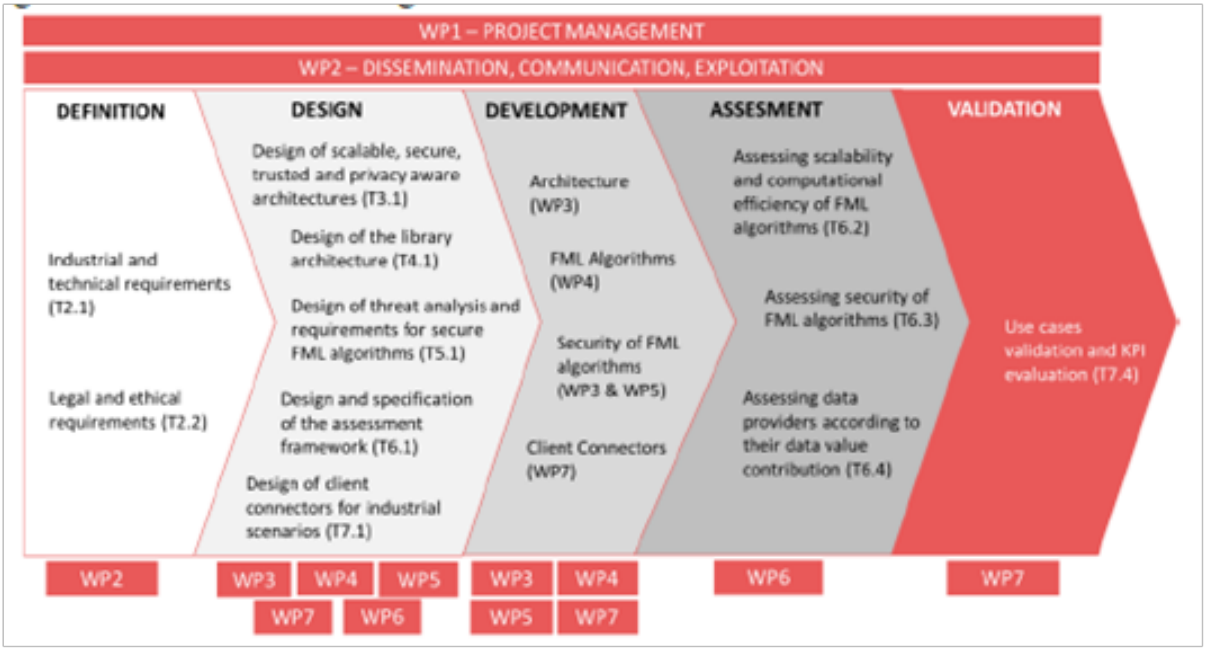

Figure 1. MUSKETEER project methodology

In particular, Stage 1 - DEFINITION will focus on the definition of the two scenarios proposed to validate the MUSKETEER platform and analysing the requirements of the different users.

In the context of MUSKETEER, in this stage, representatives of the two validation scenarios are approached to discuss the requirements. To this end, FCA and COMAU (Smart Manufacturing) and B3D and HYGEIA (Health) are directly involved as partners. FCA provide ICT methodologies and services to companies of Fiat Chrysler Group Automobiles (FCA Group) and CNH-Industrial, in the same use case COMAU provides products and technologies to meet specific manufacturing needs for industries ranging from automotive, railway and heavy industrial to renewable energy and general industry. FCA and COMAU will share the value of their data in the first use case. On the other hand, B3D provides cutting edge software technologies to improve healthcare by better extracting diagnostic data and transforming it into usable information available at the point-of-care; together with HYGEIA, a Greek Hospital composes the second use case.

These representative users have been involved in the definition process, through meetings and bilateral discussions in order to collect their requirements.

2.2 Requirements Definition

With regards to the industrial and technical requirements definition (Stage 1), an agile approach will be followed, as described in Figure 2.

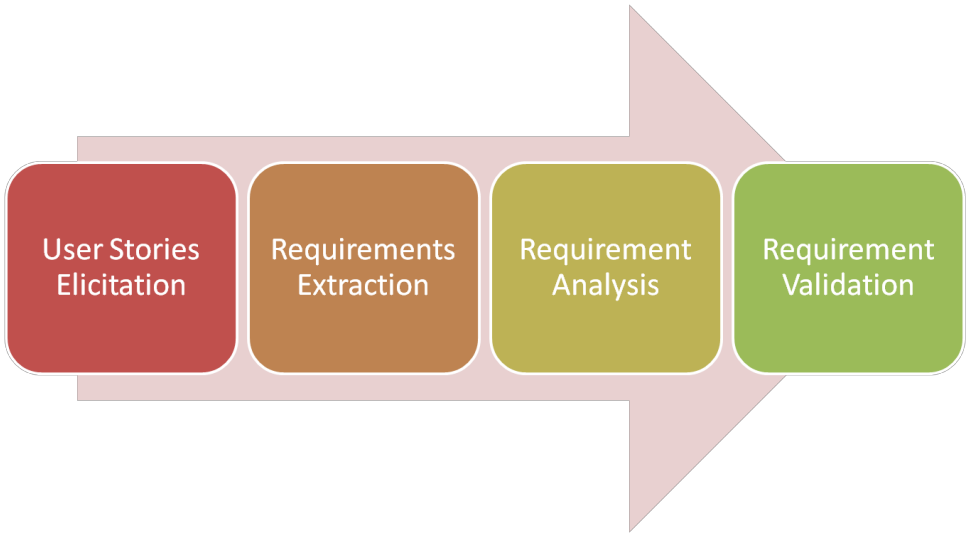

Figure 2. Requirements elicitation in MUSKETEER

In the first step, the state of research of the project topic is analysed and first high-level versions of the MUSKETEER demonstrators are described. Out of this knowledge a set of user stories are defined related to demonstrators.

A user story is a mean used in Agile software development to capture a description of a software feature from an end-user perspective. The user story describes the type of user, what they want and why. A user story is very high level and helps to create a simplified description of a requirement.

Usually, a user story provides a plain sentence to describe features of the software system to be developed. Such a sentence may lead to a reasonable work load estimation. Furthermore, the user story is used in planning meetings do enable the developer to design and implement the product features.

A user story typically owns a predefined structure:

As a <user-type (stakeholder)>, I want to <user-requirement> so that <reason>.

Then, user requirements are derived out of the user stories. They can be split into two main categories. Functional requirements define the required behaviour of the system to be build, as reported by a hypothetical observer envisioning the inputs that the product increment will accept and the outputs it will produce in response to the inputs. They are based on system objectives and respond to the critical task of ensuring the right implementation of the expected functionality in the final software. The correct and complete implementation of functional requirements will be verified by tests. Non-functional requirements define system attributes such as security, reliability, performance, maintainability, scalability and usability. Also known as system qualities, they are just as critical as functional requirements. They ensure the usability and effectiveness of the entire system. Failing to meet any of them can result in systems, that fail to satisfy business or markets or user needs.

Then, user requirements are the input for the next action, where the technical partners are translating “user” requirements more implementation-oriented requirements, which can be interpreted by system architects or developers.

During the analysis step, requirements are checked, so to avoid redundancies and inconsistencies and harmonize the level of detail. Finally, the requirements will be validated to ensure that the requirements achieve stated business objectives, meet the needs of end-user partners and are clear and understood by the developers.

3 Characterisation of the Industrial Scenarios

3.1 Smart Manufacturing

Motivation - The presence and use of robots, more generally of equipment and tools, in FCA's factories is more and more pervasive and will be more and more in the years to come. High

quality manufacturing processes require a high number of person-hours spent in the configuration of the robots and their posterior maintenance with routine inspections. However, reducing the number of inspections is a bad strategy that can reduce a plant's overall productive capacity by 5 to 20% since robots use to degrade until quality problems arise and it is necessary to stop the manufacturing plant.

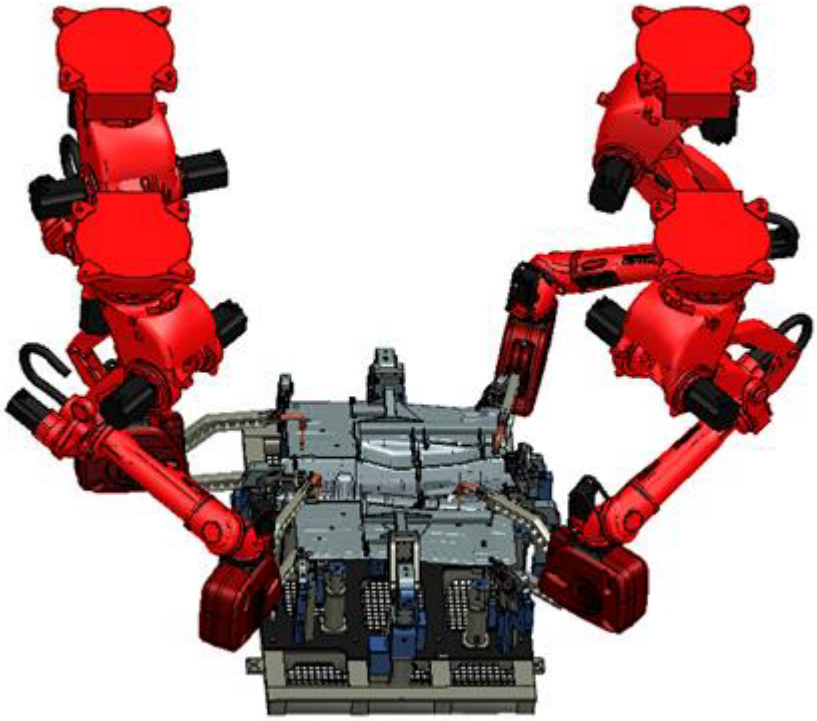

Figure 3. Welding Robot

This cost can be highly reduced with smart manufacturing thanks to the introduction of machine learning to define and update the robot settings. A predictive model, trained on historical records, can be used to alert of a possible future failure or a decrease of quality, allowing a more efficient maintenance. However, most of the times there are not enough historical records to solve these tasks collecting data from only one robot.

Since a single robot manufacturer creates instances of the same type of equipment and they can be used to perform the same operations in different sites, a potential benefit of combining the historical records of all of them can be used to improve the predictive models, with potential impacts on the product quality and plant efficiency.

Other advantages can be obtained combining historical records, not only belonging to different plants of the same company, but also to different car manufacturers. This could bring to a benefit for all the companies involved and also to the equipment supplier, which in this case is represented by COMAU.

In order to identify the effects of degraded conditions and consequent quality problems in advance of the AS IS, a possible approach is to collect and analyse all the configuration and use parameters, for example of the same class of welding guns, with the aim of searching for any correlation between the imprecision found and the conditions that generated it.

It is also necessary that the stakeholders involved share a data collection and analysis platform that is reliable and secure and that guarantees data protection.

Sharing and analysis data must favour the creation of a reference model, based on Artificial Intelligence techniques, which, appropriately fed and trained, through the use of shared data, is able to guarantee the quality for that specific operation, at desired levels and with a positive impact on the final process.

Objectives - The purpose of the use case is to collect and analyse the data related to the welding process, available in the various plants, in order to search, with the support of artificial intelligence technologies, any correlations among the variables that characterize the process so that a produced welding point will be of the expected quality.

The welding process can be represented as a decomposable dynamic system, as shown in the following figure, in Inputs, Outputs and Disturbance events.

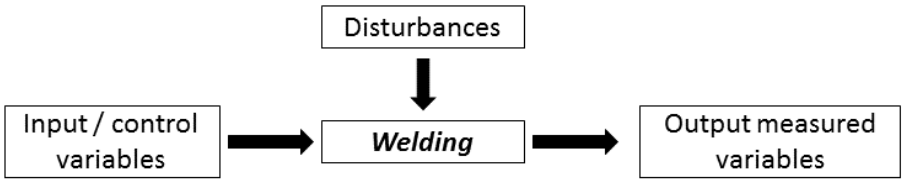

Figure 4. Welding process logical representation

As input we mean the value of the optimal parameters to be set to obtain an expected quality welding; as output we mean the measured final result, e.g. the input can be the current between the two electrodes set to obtain a right welding point; as interference factor all the boundary events that affect the expected result, like the metal sheets quality; as output the measured current effectively used by the welding gun.

In particular, as regards the inputs, the parameters involved are: welding current, running speed of the sealing element, pressure of the sealing element on the surface to be welded.

Regarding to the disturbing factors that are mainly involved in a welding process, it is necessary to take into account: mutual position of the elements to be welded, impurities of the

surface to be welded, state of wear of the sealing element and stabilization of electric current used for welding. As output, we mean: evidence of detected defects and classification of the defects detected with the methods in use (e.g. visual inspection or indirect measurement by ultrasound).

Impact - Today, poor maintenance strategies can reduce a plant's overall productive capacity between 5 and 20%. Recent studies also show that unplanned downtime costs industrial manufacturers an estimated $50 billion each year. It can be difficult to determine how often a machine should be taken offline to be serviced as well as weigh the risks of lost production time against those of a potential breakdown. Machine Learning can create predictive model to improve the quality of the manufacturing processes and to detect errors and future failures, however good predictive models require combining datasets of similar equipment in different factories. Here we can identify several of the previously defined barriers (data confidentiality, data ownership, uncertain data value, adversarial attacks) that will be avoided thanks to MUSKETEER. FCA and COMAU will share their data to identify and develop methods and techniques that allow collection and analysis of the data related to the welding process, available in the various plants, in order to search, with the support of artificial intelligence technologies, any correlations among the variables that characterize the process so that a produced welding point will be of the expected quality and the remaining time to loss of quality. The availability of the correlations sought enables the improvement of the welding process with positive impacts both on the quality of the welding process and on the final product associated with it. In particular, the sharing and analysis of data could generate a distribution curve of the probability of error, in qualitative terms, stabilized within a set interval and such that the impact on the quality of the final product is acceptable. Moreover, the estimation of the remaining time to loss of quality can improve the maintenance organization, avoiding replacements of still functioning pieces and, on the other hand, avoiding unexpected failures.

This will have a direct impact in the reduction of the manufacturing process due to: (1) An improvement of the welding process with positive impacts both on the quality of the welding process and on the final product associated with it; (2) A reduction in the welding gun maintenance cost.

Welding process - Welding is the process by which two pieces of metal can be joined together thanks to a fusion of the layers. The welding gun, shown in the following figure, is composed by:

- Two mechanical arms, one fixed and the other which can move;
- A linear motor which allows the arm movement;

- A copper electrode at the end of each arm, which is in contact with the metal sheets to weld;
- A water cooling system;
- A welding tray which is the controller of the current supplied for the welding.

Figure 5. Welding gun image

In general, the number of metal sheets to weld varies from 2 to 3. The supplied current, the time spent on the welding process and the pressure applied by the arms on the metal sheets strictly depend on the number of layers and on their thickness.

The current is supplied by the welding tray and flows through the arms up to the pieces of metal to melt.

The spot-welding time cycle (Figure 6) in detail is characterized by four time-measurements: squeeze time, weld time, hold time and off time. The squeeze time represents time between pressure application and weld; the weld time represents the weld time in cycles; the hold time represents the time the pressure is maintained after weld operation is completed and off time the time in which electrodes are separated to enable the next spot.

During each welding point the electrodes are subjected to a degradation and they get dirtier. This cause a loss of quality in following welding points. For this reason, after a predefined number of points, the electrodes undergo to a dressing process, which consists on a small material removal. After some removals the electrode has to be changed. Welding data contain information of these processes by means of counter variables.

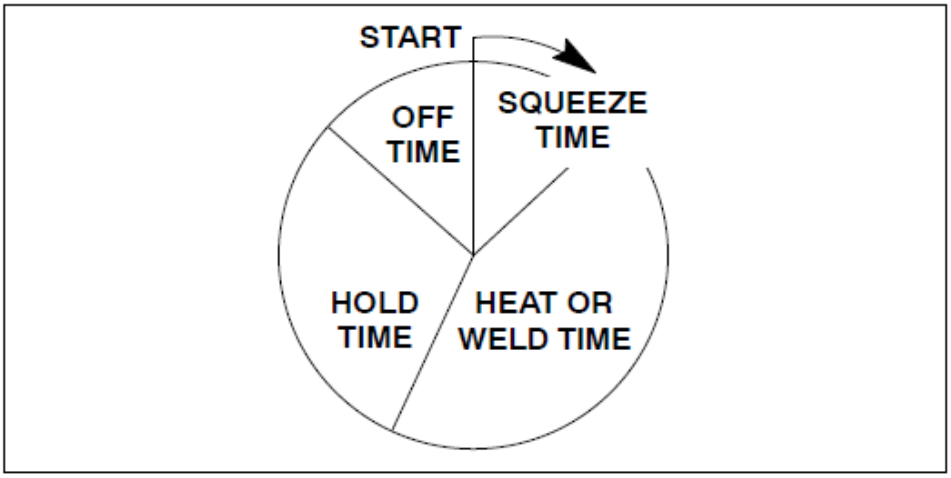

Figure 6. Spot-welding time cycle

In general, RSW is based on four major factors, which most describe the welding process:

- Amount of current that passes through the “work piece” [kA];
- Time in which the current flows through the “work piece” [s];
- Pressure that electrodes apply on the “work piece” [daN];
- The area of the electrode tip contacts with “work piece”.

3.1.1 User Stories

The users foreseen in this scenario are domain experts, data scientists, developers, in-line production operators, production managers.

These users have different motivations and requirements that are presented as user stories.

Table 1 Smart manufacturing user stories

User Story ID	As a/an	I want to	so that
MUS01	Data scientist (FCA, COMAU)	Prepare welding gun data	I can train and test ML models
MUS02	Data scientist (COMAU)	Have data from different car manufacturer sources but comparable	I can obtain significant results

User Story ID	As a/an	I want to	so that
MUS03	Data scientist (FCA)	Have data from different production plant sources but comparable	I can obtain useful results to evaluate welding gun behaviour in terms of welding quality
MUS04	Data scientist (FCA, COMAU)	Select validated features or ML models	I can share learning with other institutions
MUS05	Data scientist (FCA, COMAU)	Select features or ML models shared by other institutions	I can improve my ML models
MUS06	Data scientist (FCA, COMAU)	Have data collected for a long time	I can analyse the decay over time
MUS07	Developer	Receive trained and validated ML models in executable application	I can easily integrate the ML models in the pre-processing pipeline
MUS08	In-line production operator (FCA)	See some alerts if there will be some failures.	I can warn my manager if the alert is worrying
MUS09	Engineer (COMAU)	Receive features of most common malfunction.	I can develop a better version of welding gun or robot controller software
MUS10	Production manager (FCA)	Become aware if some welding guns in my production lines will have some malfunction visualising a dashboard.	I can call a maintainer in advance and organize the maintenance when the plant is not in production.
MUS11	Maintainer (FCA, COMAU)	Know in advance if some failures will occur in a welding gun.	Organize my visit without urgency and acquire knowledge on consumables (spare parts) lifecycles
MUS12	Maintainer (FCA, COMAU)	Know in advance which can be the problems in the welding gun.	I can bring with me the useful spare parts.

User Story ID	As a/an	I want to	so that
MUS13	Engineer (COMAU, FCA)	Receive features of most common malfunction.	I can define my needs and the requirements of a task
MUS14	ICT specialist (COMAU, FCA)	Receive requests of grants for visualise or do active actions on the platform by my colleagues.	I can enable my colleagues to visualise platform parts or to do active actions
MUS15	ICT specialist (COMAU, FCA)	Preserve privacy of my company	I don't want to share raw data without privacy pre-serving methods
MUS16	Data scientist (FCA, COMAU)	Pre-process my data in a validated way	I can use a specific pre-processing pipeline for my data

3.1.2 Data Characterisation

The starting dataset consists of all the data extrapolated from files in which the chronological sequence of the operations, carried out during the welding process, is recorded. These data, that could be collected only if a specific software has been enabled on the welding machine, contain variable presented in the table below:

Table 2 Welding data characterization

Variable ID	Variable description	Variable type	Measure unit
DEPCOD	Department code	Text	-
SPOTNUM	Welding point code number (univocal for each point of the cycle)	Text or Int	-
PRGNUM	Welding program number	Int	-
WELMOD	Program execution mode (Normal, Monitor, Corr. Cost)	Txt	-
WFORC1	Set value of the electrodes force	Int	daN
CURMOD	Check current mode (Switch off, primary current, secondary current)	Txt	-
REGMOD	Check it quality regulator mode (on-off)	Bool	-
CLSMOD	Check it quality classifier mode (on-off)	Bool	-

Variable ID	Variable description	Variable type	Measure unit
WELCNT	Welding points counter	Int	-
WDRSCNT	Welding dressing counter	Int	-
CURRENT	Current	Int	kA
WPWR	Percentage value inverter utilization	Int	%
WELE	Measured value of the energy in the welding point		J
SPLIDX	Welding spray index		ms
TOLDINF	Diagnostic lower tolerance percentage of the current	Int	%
TOLDSUP	Diagnostic lower tolerance percentage of the current	Int	%
WELTIME1_S	Set welding time phase 1	Int	ms
WELTIME1_M	Measured welding time phase 1	Int	ms
WELCUR1_S	Set welding current phase 1	Float	kA
WELCUR1_M	Measured welding current phase 1	Float	kA
WELTIME2_S	Set welding time phase 2	Int	ms
WELTIME2_M	Measured welding time phase 2	Int	ms
WELCUR2_S	Set welding current phase 2	Float	kA
WELCUR2_M	Measured welding current phase 2	Float	kA
WELTIME3_S	Set welding time phase 3	Int	ms
WELTIME3_M	Measured welding time phase 3	Int	ms
WELCUR3_S	Set welding current phase 3	Float	kA
WELCUR3_M	Measured welding current phase 3	Float	kA

To sum up, the dataset contains three categories of data: (a) nominal setting values to obtain a welding of adequate quality, (b) values recorded during the punctual execution of a specific welding, (c) two counter variables that record the number of welding points done after the last electrode dressing and the total number of electrode dressing made.

For each point of a welding cycle the mentioned data are recorded and saved in a file. These files, that therefore contain the data of all points of the cycle over time, are stored in a server

located on premise in the production plant. In order to manage the quantity of recorded files they are periodically moved on a secure dedicated storing area.

The built dataset is the training pattern of the artificial intelligence model that will have to search for any correlations between the "imprecise" welding and the value of the parameters indicated at points (a) and (b) plus the disturbance factors introduced into the system by the boundary conditions in which it is located.

Notice that if we move the focus at a higher level, we see that the welding data that have just described are part of a more complex architecture.

In particular the data related to manufacturing scenario give a representation of different data sources that may provide useful information for the spot-welding operation. The specification of the welding gun and current setup configuration provide nominal values which characterize the parameters and the machinery state. The data provided by the supplier enable, besides, the building of theoretical preventive maintenance to be applied if the declared performances coincide with expectations. The welding domain experts explicit the relationship between weld issues and causes and the spot-welding defects assessments can be addressed through machine learning in order to refine the relationship, to assure products quality and to study the welding gun degradation time.

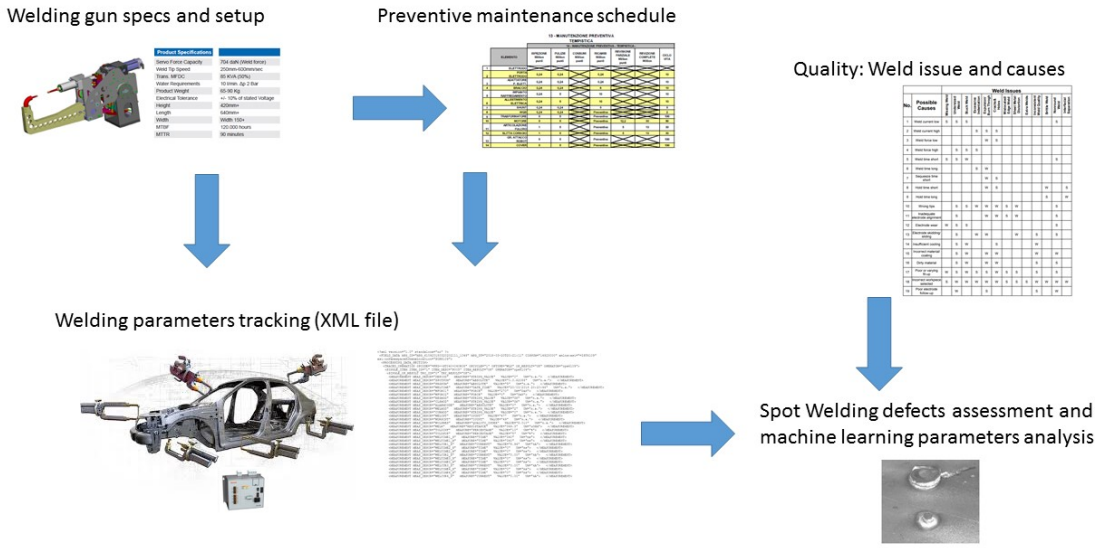

Figure 7. Smart Manufacturing local data collection

3.1.3 The Data Flow

The data flow that will validate the MUSKETEER Platform in Manufacturing environment will generate, collect in files and stored welding data on premise in different production plants of FCA. The computation of a machine learning model using these data will also be run locally and then this model will be sent to the MUSKETEER Platform instance based in COMAU. At this point different models will be combined in order to enrich shared knowledge, and then a better model will be sent to the plants.

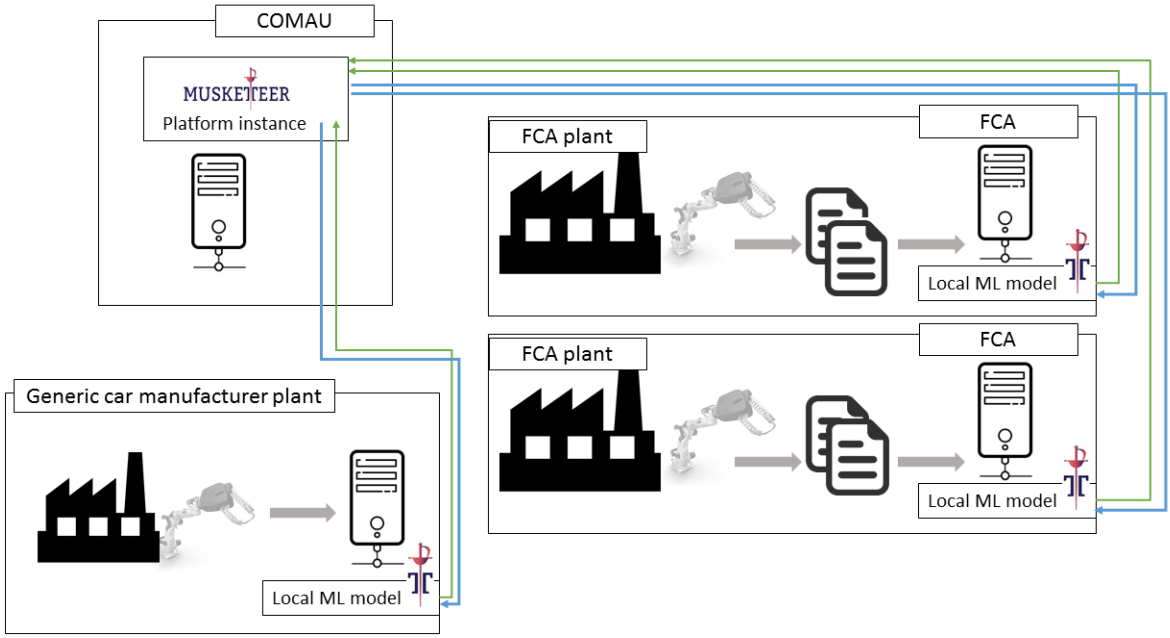

Figure 8. Smart Manufacturing architecture

3.1.4 Privacy Operation Modes

Privacy and security concerns to overcome - The main privacy problem is data confidentiality. The historical records contain information about the industrial processes of a company and used solutions. Any information leakage can potentially reveal industrial secrets about internal manufacturing processes and the problems that production teams has to face with in the plants. This information can give competitive advantage to OEM competitors and cause damages to brand. That's why we will use IDSA concepts and models to ensure confidentiality and privacy protection to the IDS ecosystem stakeholders.

In addition, a factory may be hit by cyber-attacks. A data poisoning attack can produce a useless predictive model and the malfunction of the production plant. An attack can lead to a false alarm of a possible future failure and a maintenance cost increasing.

The most suitable POM for the industrial use case, taking into account the privacy requirement of organizations involved (FCA and COMAU), at this stage of knowledge, is PORTHOS presented in Figure 9.

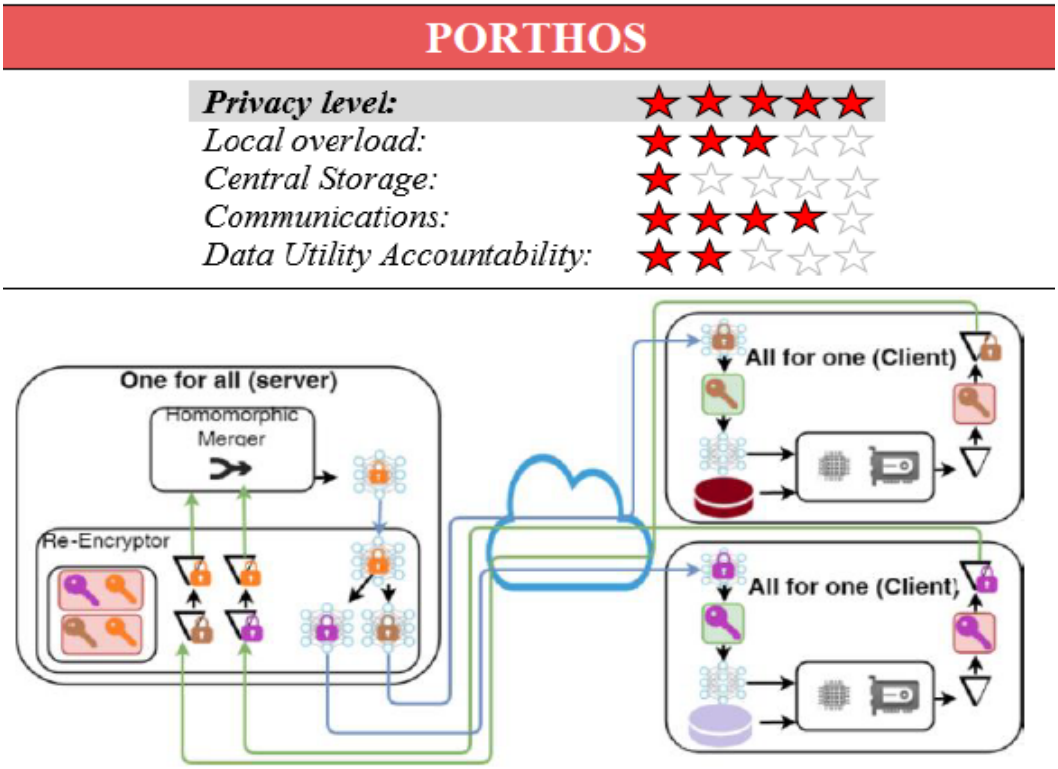

Figure 9. POM PORTHOS

3.2 Health – Medical Imaging

Motivation - Health data is a very special type of personal data that encompasses an extreme value for the person itself, considering its own health and wellbeing, and for the healthcare practitioners who should decide on the correct diagnosis and care pathways to achieve the best patient outcomes. Health data is also extremely important to the research, development and validation of new technologies, procedures and care pathways to improve the diagnosis, prognosis and treatment of diseases.

The recent years have shown important advances in Artificial Intelligence, enabled by cloud-computing and big-data collections, with application in many different fields, and also with strong promises in the health care sector. One key element for improving AI algorithms and its results is gathering large amounts of good-quality data. In the health care sector, mainly for security and privacy reasons, but also due to some lack of interoperability and standardisation, it has been difficult to concentrate large amounts of quality data for the development

of AI methodologies. Biobanks are vital source of information for fundamental and translational biomedical research aimed at the development of better predictive, preventive, personalised and participatory health care [31]. Although 70% of world biobanks are located in Europe, until recently, imaging data coming from sources such as magnetic resonance imaging (MRI) or computed tomography (CT) were not included in such biobanks [32]. Projects have been launched to acquire large repositories of image data, but in 80% of cases the access to imaging biobanks is restricted to research and clinical reference.

Multi-tenant and multi-datacentre cloud solutions for medical imaging management, analysis and reporting, have been used in clinical practice for radiology and tele-radiology for a few years. They have been used by public hospitals to organise networked, collaborative reporting services, and by private practices to improve the productivity on large distributed groups and on small clinics. Vast amounts of medical imaging data are collected and reported using these cloud solutions, but each organisation accesses only its own data. Thanks to MUSKETEER this limitation will be surpassed.

The pressure for productivity is increasing due to the lack of Radiologists and the growing demand for medical imaging services. Key driving factors are the rise in prevalence of chronic diseases, technological advancements in diagnostic imaging modalities, increasing number of imaging procedures, rising awareness among the patients about early diagnosis of clinical disorders and rise in base of aging population. In addition, increasing demand from emerging countries, improved government funding towards chronic disorders, increasing investment in public and private organizations, and increasing disposable income among the population will further expected to drive the market in the coming years.

The European Union, the US and many other countries have been focusing their public health policies and research efforts on personalised medicine and evidence based clinical pathways to improve patient outcomes and effectiveness of care. The solution lies in providing powerful tools to support the radiologists to take faster and more accurate decisions for diagnosis and prognosis. As some research projects have been indicating, AI and Deep Learning (DL) are the disruptive technologies that will enable develop these powerful tools leading to improvements to the clinical protocol pathways and conducting to better efficiencies and better patients' outcomes.

Radiology is moving toward a future in which radiologists, guided by artificial intelligence, will be able to work more closely with clinicians to provide precise therapies that offer patients an improved quality of life (according to a series of speakers at the opening press conference of the European Congress of Radiology ECR2019).

Objectives - This use case intends to demonstrate the application of the Artificial Intelligence methodologies and technologies developed on other WPs, enabling access to vast amounts of

distributed medical imaging data to train and improve the learning algorithms, providing powerful tools to improve clinical practice.

Being such a vast area, with several imaging modalities applying to different human body parts to analyse distinct conditions, we shall restrict the demonstration to one specific type of study. The main objective will be the training of AI algorithms for support the detection of prostate cancer.

We keep open the decision to include another relevant type of study to be addressed under MUSKETEER proposal, such as liver or colon cancer. The decision will be taken depending on different variable like, re use of machine learning algorithms, interest of the medical expert in HYGEIA, impact to the health sector, etc.

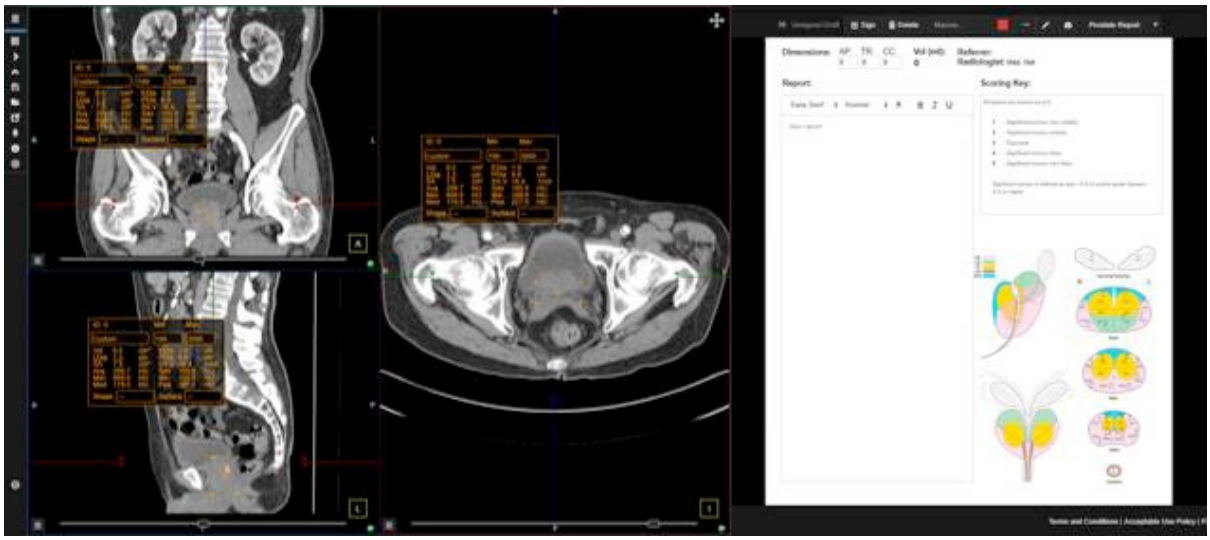

Figure 10. Prostate image visualisation and report

Impact - B3D and HYGEIA will take a huge advantage of MUSKETEER developments to demonstrate the application of the Artificial Intelligence methodologies and technologies enabling access to vast amounts of distributed medical imaging data to train and improve the learning algorithms, providing powerful tools to improve clinical practice.

Being such a vast area, with several imaging modalities applying to different human body parts to analyse distinct conditions, we shall restrict the demonstration to one specific type of study. The main objective will be the training of AI algorithms for support the detection of prostate cancer. Since it is really hard to collect medical records, the benefit to collaborate sharing datasets to improve the predictive models to aid in the medical diagnosis is clear. The main barriers to be avoided with MUSKETEER are data localization, information leakage, standardisation and adversarial attacks. This project can solve these barriers.

The expected impacts of this use case are several:

1. Improve accuracy of AI algorithms by sharing knowledge from distinct organisations and data repositories, supporting cooperation keeping security and privacy of health data;
2. More accurate clinical decision support tools for diagnosis and prognosis of diseases, avoiding invasive procedures and conducting to better patient outcomes;
3. Faster decision support tools, enabling shorter turn-around-times, increasing productivity of services and more studies and patients diagnosed;
4. Faster and more accurate clinical decision support tools for diagnosis and prognosis saving lives in emergency cases;
5. Enable the growth of the level of research in medical imaging AI tools supported by distributed data repositories;
6. Enable clinical practices to access medical imaging AI tools with gains of productivity and better patient outcomes;
7. Improve Biotronics3D commercial offer, enabling partners to access its market.

3.2.1 User Stories

The users foreseen in this scenario are domain experts, medical doctors and radiologists, data scientists, developers, and administrators. These users have different motivations and requirements that are presented as user stories.

Table 3 Health – medical imaging user stories

User Story ID	As a/an	I want to	so that
HUS01	MD or Radiologist	Quickly observe the regions of interest in an imaging study with segmentation	I can avoid many images without interest and focus on those that present any abnormality
HUS02	MD or Radiologist	Get qualitative and quantitative information on the regions of interest	I can take the correct decision on the final diagnosis and improve patient outcomes
HUS03	MD or Radiologist	Get information about AI models' performance	I can opt for what model to use

User Story ID	As a/an	I want to	so that
		(Sensitivity, specificity, confidence intervals, ROC…)	
HUS04	Data scientist	Select cases from imaging and diagnosis report database	I can train and test AI models
HUS05	Data scientist	Select validated features or AI models	I can share learning with other institutions
HUS06	Data scientist	Select features or AI models shared by other institutions	I can improve my AI models
HUS07	Developer	Receive trained and validated AI models in executable application	I can easily integrate the AI models in the pre-processing pipeline
HUS08	Administrator	Define acceptance criteria for models' accuracy	I can share and receive good quality information
HUS09	Administrator	Define users' permissions	Each user accesses the right modules.
HUS10	Administrator	Monitor data exchange with 3rd parties	I can assure the anonymity of shared data, the trustworthiness of 3rd party sources, and compliance with GDPR

3.2.2 Data Characterisation

Biotronics3D provides a cloud-based Medical Imaging platform, 3Dnet™, for storage, retrieval, conditioning, fusion, analysis, presentation, interaction and reporting studies across medical imaging industry. 3Dnet Medical cloud collects and manages vast amounts of data for distributed services of Radiology and Teleradiology, using DICOM and HL7 standards, into its secure cloud infrastructure, providing advanced visualization techniques to the Radiologists, allowing to make the right decisions and report their diagnosis, anytime and anywhere.

Hygeia is a reference organization for health care services in Greece, being the first hospital throughout Europe to carry out implantation of radioactive particles in prostate cancer.

Hygeia has its own datacentre and uses 3Dnet Medical software to manage the medical imaging data.

Considering the objective in this project is the training of AI algorithms for support the detection of prostate cancer, follows the characterisation of prostate cancer imaging and reporting data.

In terms of input data for system training reasons, Hygeia will draw all pelvis MRI exams as well as multiparametric MRI exams for male patients. For each exam, an assessment sheet (in pdf format) is attached where lesions (regions of abnormality) are shown in images with PI-RADS score for each lesion (grading from I to V). Available histopathology reports (written as text in Greek language) from these exams will also be gathered.

All the above data is delivered as input to the software with main objective the training of AI algorithms and the potential to get precise details as an output on lesions localization and corresponding PI-RADS score.

The PI-RADS is described in the document Prostate Imaging Reporting & Data System - PI-RADS 2015 Version 2, of the American College of Radiology (ACR). A graphic presentation of the scoring system is provided by the Radiology Assistant site.

The PI-RADS scoring system has the following grades:

- PI-RADS 1: Very low (clinically significant cancer highly unlikely)
- PI-RADS 2: Low (clinically significant cancer unlikely)
- PI-RADS 3: Intermediate (clinically significant cancer equivocal)
- PI-RADS 4: High (clinically significant cancer likely)
- PI-RADS 5: Very high (clinically significant cancer highly likely)

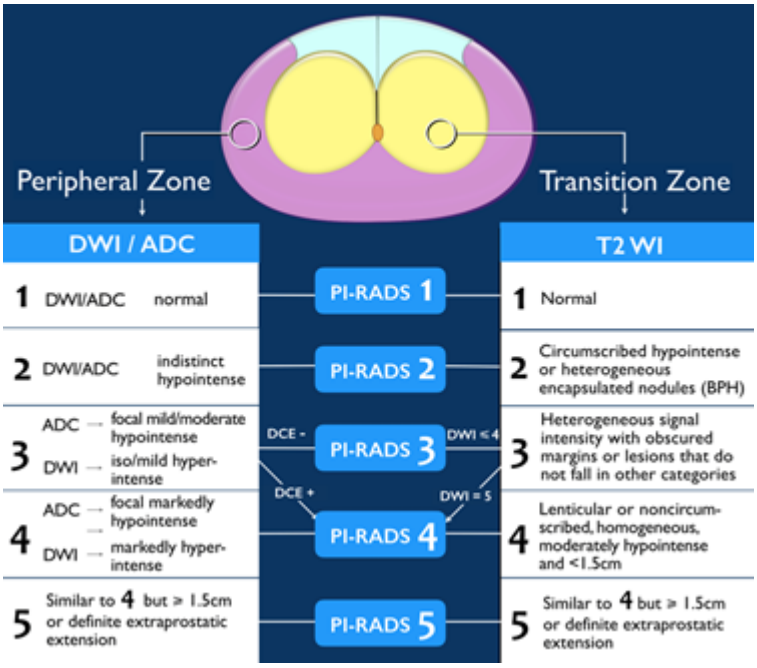

Figure 11. PI-RADS assessment (from radiologyassistant.nl)

The image segmentation should follow the sector map used in the PI-RADS version 2 which employs 39 sectors (12 in the base, 12 in the midportion, 12 in the apex of the prostate, 2 seminal vesicles and 1 urethral sphincter).

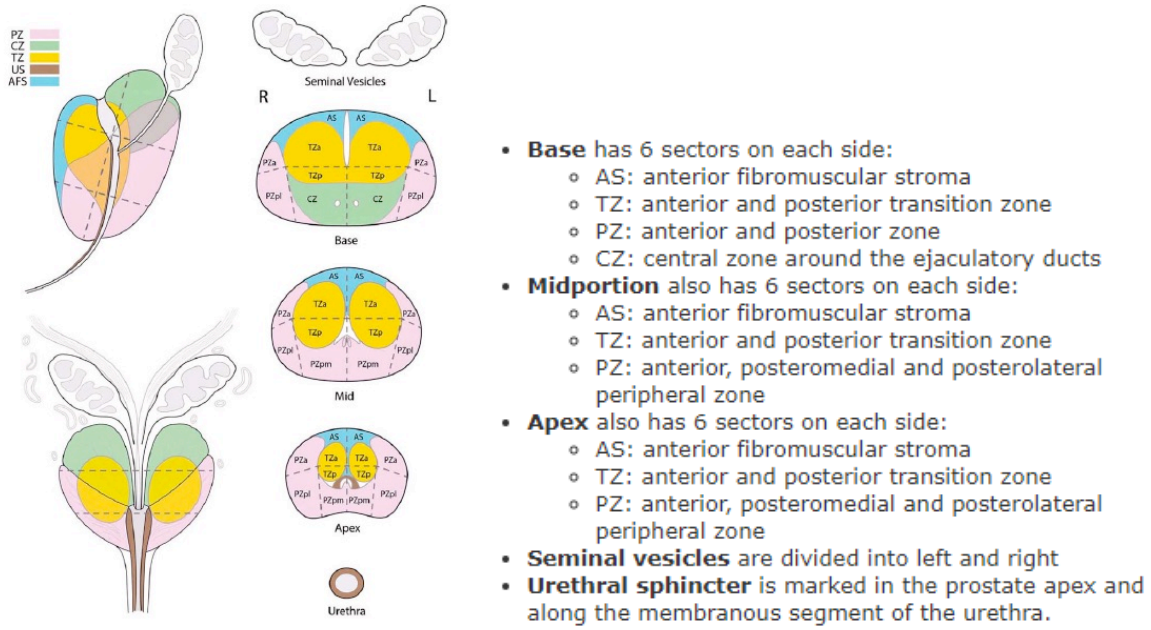

Figure 12. PI-RADS segmentation (from PI-RADS™ v2)

The radiologyassistant.nl website presents an animation video about the zonal and sector anatomy of the prostate (https://youtu.be/cWTJsJFhjA4).

Volume of existing data in number of studies and reports (data available for training):

Table 4 Available data

Exam	Period	#Exams	#Reports
MRI Pelvis Male	2018	383	370
MRI Pelvis Male	2017	298	290
MRI Pelvis Male	2016	283	280
MRI Pelvis Male	2015	354	310
MRI Pelvis Male	2014	368	300
MRI Pelvis Male	2013	48	0
MRI Pelvis Male	2012	27	0
MRI Pelvis Male	2011	32	0
MRI Pelvis Male	2010	26	0
Multiparametric MRI of Prostate	2010-2019	346	290

The frequency of new data is presented on the following table.

Table 5 Average frequency of data

Modality	#Studies	#Reports	#Gold Standard
MRI	40/month	40/month	

Storage of data:

MRI studies are stored in the 3Dnet PACS imaging database in DICOM format.

Reports are also stored in 3Dnet system or in a separate RIS system, communicating using HL7 standard.

Biopsy reports are stored in the Laboratory Information System and can be uploaded on 3Dnet using HL7 standard.

Additional data sources:

There are open repositories with public access data that may be used for training. Biotronics3D can create a special organisation in 3Dnet cloud to hold data from these repositories and provide access to MUSKETEER partners for training.

The Cancer Image Archive (TCIA) is a service which de-identifies and hosts a large archive of medical images of cancer accessible for public download. The data are organized as "Collections", typically patients related by a common disease (e.g. prostate cancer), image modality (MRI, CT, etc) or research focus. DICOM is the primary file format used by TCIA for image storage. Supporting data related to the images such as patient outcomes, treatment details, genomics, pathology, and expert analyses are also provided when available. TCIA have 7 collections related with prostate, 6 of them containing MR studies of +500 subjects.

Zenodo and OpenAire contain several collections of prostate MRIs with open access as the two following samples demonstrate:

"Annotated MRI and ultrasound volume images of the prostate", March 26, 2015, DOI: 10.5281/zenodo.16396 (https://zenodo.org/record/16396#.XIZSNyj7Tcs).

"Original multi-parametric MRI images of prostate", make available by I2CVB: http://i2cvb.github.io/. Publication date: October 20, 2016, DOI: 10.5281/zenodo.162231.

3.2.3 The Data Flow

Medical imaging studies are obtained with scanners and uploaded to the 3Dnet PACS system using DICOM standard, which defines data structures and communication protocols. Medical images are analysed in 3Dnet by radiologists and/or other medical specialists that produce a medical report and may annotate the images. Additional diagnosis information, like biopsy report, can be uploaded on 3Dnet from Laboratory Information System using HL7 standard.

A local MUSKETEER module should receive data for training the local ML models and allow users to select which external models they want to combine to improve learning models. The user should be able to select the ML model to use in production and the models to share with other institutions.

After defining a production ML model, when a new study arrives in 3Dnet it can be sent automatically to the local MUSKETEER module to run the production model on that study and send back results. When radiologists and/or other medical specialists open the study in 3Dnet, they obtain qualitative and quantitative information on the regions of interest and performance metrics about the production ML model.

At central level, the MUSKETEER platform instance enables the sharing of models between institutions and the combination of different models in order to enrich the overall knowledge.

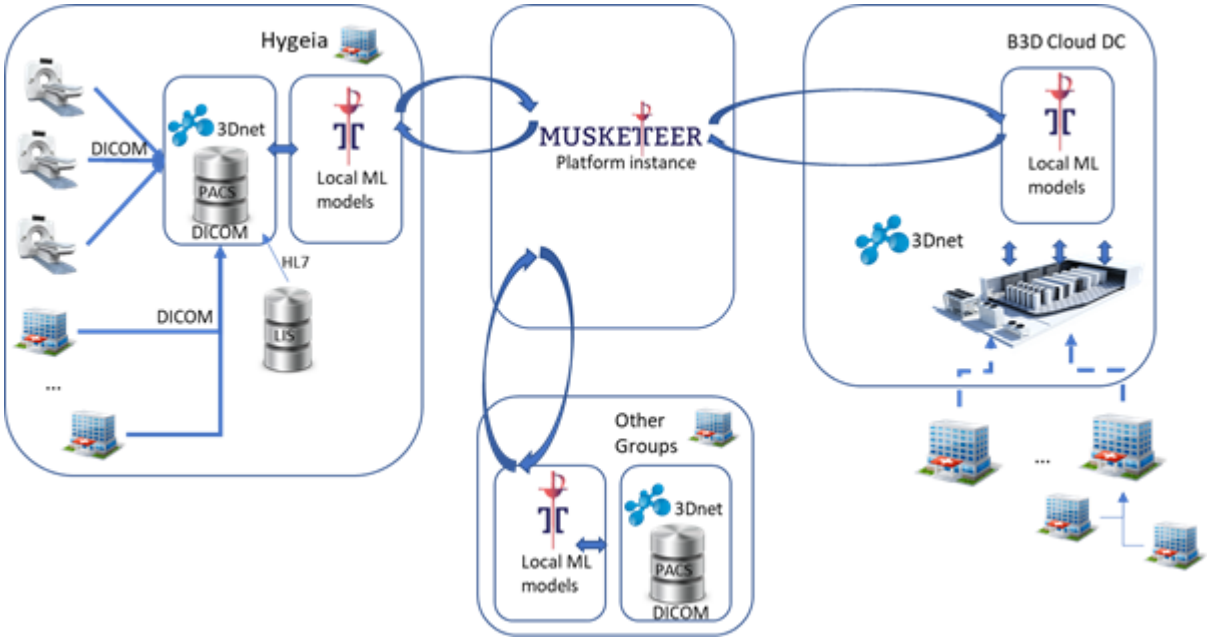

Figure 13. Medical Imaging data flow

3.2.4 Privacy Operation Modes

The main privacy problem in the health care scenario is the security and privacy of personal data.

Machine learning algorithms can process health records to create predictive models capable to help in the medical diagnosis, these types of datasets are very valuable for research purposes. For a single hospital it is very complicate to collect a dataset large enough to create a complex predictive model. For that reason, the benefit, in terms of predictive model accuracy, of combining datasets of different hospitals is very clear. However, having explicit consent of a patient to use his/her health records does not guarantee the protection of security and privacy and when two or more different research groups have explicit permission to use a health dataset, different barriers arise.

Data Localization barriers: To create a predictive model currently is necessary to place the dataset in a single place (same local computer or in the same cloud computing cluster). However, data localisation among different countries stems from legal rules that dictate the localisation of data for its storage or processing. Such requirements restrict the free flow of data between regions or countries.

Information Leakage barrier: Even signing a non-disclosure agreement, digital information can be easily copied and redistributed. Giving directly access to other's datasets open a door to personal data robbery. This entails severe fines for hospitals and personal damage if personal information is revealed.

Standardisation Barrier: When different hospitals create a dataset (e.g. medical images) using different medical devices (of different device manufacturers or with different calibration), then every hospital can use completely different measuring units and data standardization takes special importance.

Data Untrustworthiness Barrier: Groups may distrust the other's datasets since some partners can make a data poisoning attack in order to slow down the research of other group in a specific field.

MUSKETEER allow machine learning over datasets allocated in different locations (thus removing the data localization barrier) where the privacy preserving analytics remove any chance of information leakage and with mechanisms to provide standardisation among different partners (based on IDSA concepts and Reference Architecture Model). In addition, the adversarial attack detection and mitigation strategies will be capable to detect data poisoning attacks and alert the other hospitals.

The most suitable POM for this use case, taking into account the privacy requirement of organizations involved, are PORTHOS, showed in Figure 9, and RICHELIEU, presented in Figure 14.

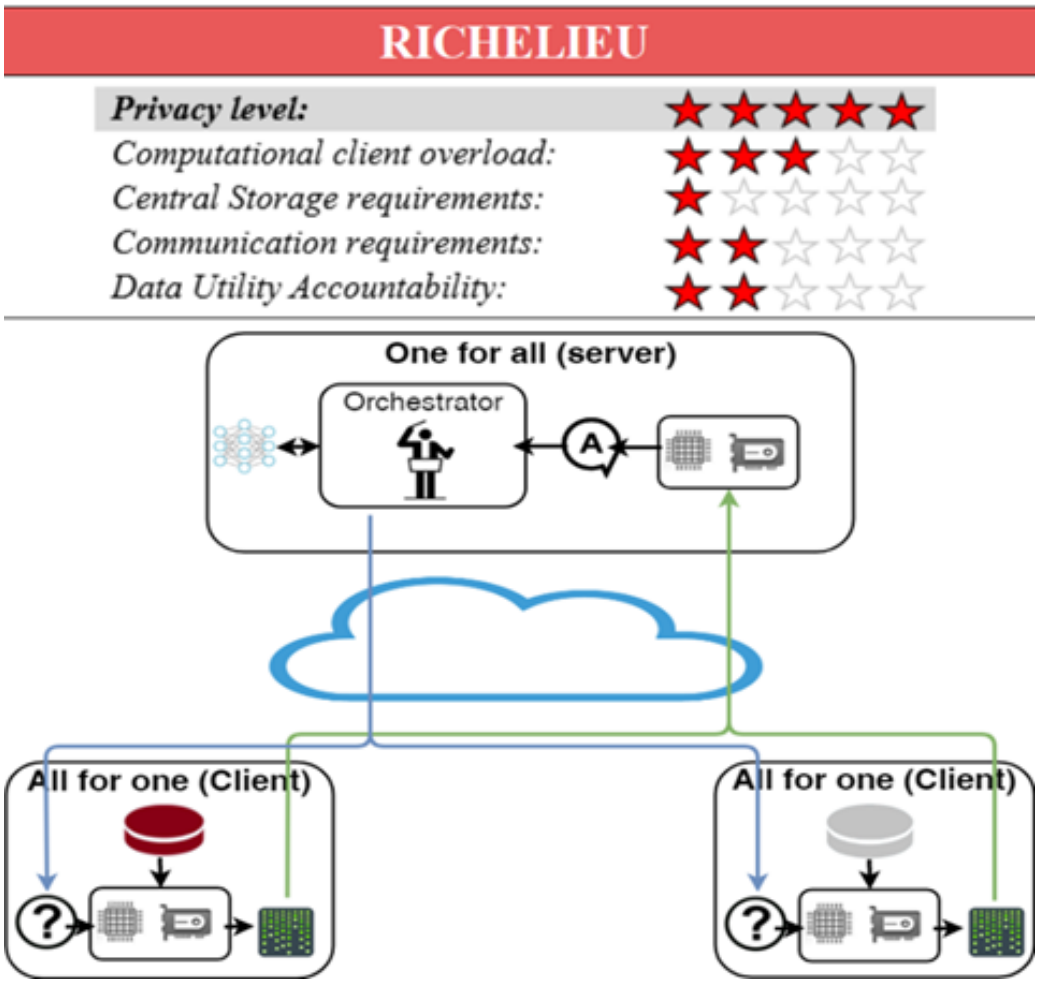

Figure 14. POM RICHELIEU

3.3 MUSKETEER User Goals

In this section, a first iteration to the definition of goals at a project level is presented, based on the user stories of the industrial scenarios. They could be considered a first input for the task T2.3 where the Goal-Question-Metric (GQM) methodology is to be considered to define the MUSKETEER evaluation framework and overall evaluation approach that will be implemented in WP7 according to the different MUSKETEER use cases (see the deliverable D2.3 (M6) for more details).

The Goal-Question-Metric (GQM) methodology used in agile environments, allows identifying and further refining of explicit measurement goals, eliciting a set of questions to achieve each goal and identifying metrics to answer those questions:

- Goals define what the project wants to improve;
- Questions refine each goal to a more quantifiable way;
- Metrics indicate the metrics required to answer each question.

From the user stories it is possible to identify the goals for each of the industrial scenarios. Although the goals are different for each scenario, they follow the same general ideas:

- Improve ML models for detection/prediction of critical conditions (failure / lesion);
- Obtain fast and accurate ML models;
- Share learning to improve ML models;
- Obtain trained and validated ML models ready to deploy;
- Improve security and confidentiality of data.

4 Specification of Requirements

This section presents the detailed specifications of user and technical requirements for the MUSKETEER platform based on the deep analysis of the industrial scenarios. First, the roles of users in MUSKETEER platform are described and mapped to the industrial users identified in the industrial scenarios. User requirements are specified based on the user stories description and categorised into functional and non-functional requirements. Finally, the technical requirements translate the user requirements into implementation-oriented requirements.

4.1 User Roles

Table 6 MUSKETEER User roles

User role	Description
General user	A general user has access to the platform through an account. This is the most general user role of the platform, with visualisation grants.
Technical user	A technical user is a general user and has the possibility of active actions on dataset and tasks based on his grants.
Task creator	A task creator user is a technical user and has the possibility of create a task. They are the owners of the task and have specific privileges, like inviting / accepting contributions by other users to the task.
Task member	A task member user is a technical user that participate of a task. They are not the initiators of the task, but can contribute data / model updates and can benefit from the resulting trained model (depending on the POM).
Group owner	A group owner is a technical user and his/her role facilitates the sharing of data / models between members of the same organization or groups of organisations.
Researcher	A researcher is a technical user with the specific role of benchmarking the performance of the platform. As such, they need to be able to run synthetic tasks on artificial data, involving multiple artificial users.
Platform admin	A platform admin is a technical user and has administration capacities on the platform, he/she has the possibility of active action on users grants.

4.2 Industrial Users' Roles in MUSKETEER Platform

The mapping between the users from the industrial scenarios and their roles in the MUSKETEER platform is presented in Table 7. Most industrial users only have general user roles in the platform because they only need to know the performance of the selected ML models developed and trained in MUSKETEER platform by data scientists and deployed locally by developers in the processing pipeline.

Table

Table 7 Mapping between industrial users' roles and MUSKETEER user roles

Users X Roles	General user	Tech-nical user	Task creator	Task member	Group owner	Re-searcher	Platform admin
Data scientist	X	X	X	X	X	X	
Developer	X	X		X			
Production operator	X						
Engineer	X	X					
Production manager	X						
Maintainer	X						
MD or radiologist	X						
Administrator	X	X			X		X

4.3 MUSKETEER user requirements

The following tables list the derived requirements for the classes

- Functional requirements,
- Non-functional requirements.

4.3.1 Functional Requirements

Table

Table 8 Functional requirements

ID	Description of the requirement	Category	Related User Sto-ries	Platform role
FR001	Login with username and password.	Accounting	All	General user
FR002	Change password and update profile information.	Accounting	All	General user

FR003	Add and remove information about datasets (or the datasets themselves) of user ownership. Manage its visibility to other users or groups.	Configuration	MUS01, MUS02, MUS03, MUS06, HUS04, HUS10	Technical user
FR004	Manage own visibility.	Configuration	All	General user
FR005	Browse other users.	Usage	All	General user
FR006	Browse datasets owned by other users.	Usage	MUS01, MUS02, MUS03, MUS06, HUS03, HUS06, HUS10	General user
FR007	List all the tasks that have been created.	Usage	MUS01, MUS04, MUS05, MUS07, HUS06, HUS07, HUS08, HUS10	Technical user
FR008	Search tasks using several search options (description, status, sort of data).	Usage	MUS01, MUS04, MUS05, MUS07, HUS06, HUS07, HUS08, HUS10	Technical user
FR009	See summary statistics of all the tasks that have been created (total number of tasks, number of participants, compensation / data value).	Usage	All	General user
FR010	View details of tasks: high-level description, which model, what sort of data, what is the status of the training, how many participants etc.	Usage	MUS01, MUS02, MUS03, MUS04, MUS05, MUS06, MUS07, HUS06, HUS07, HUS08, HUS10	Technical user
FR011	Search models using several search options (description, sort of data, performance, accuracy).	Usage	MUS01, MUS02, MUS03, MUS04, MUS05, MUS06, MUS07, HUS06, HUS07, HUS08, HUS10	Technical user

FR012	Join a task that has already been created and that accepts new participants.	Usage	MUS01, MUS04, MUS05, MUS06, MUS07, HUS06	Technical user
FR013	Download trained machine learning models (or part of) if permitted.	Usage	MUS01, MUS05, MUS07, HUS06, HUS07, HUS08, HUS10	Technical user
FR014	Request download permissions for trained ML models.	Usage	MUS01, MUS05, MUS07, HUS06, HUS07	Technical user
FR015	Pay for ML model download permissions if required.	Usage	MUS01, MUS05, MUS07, HUS06, HUS07	Technical user
FR016	Initiate to train a ML task: define the task, select participants or datasets.	Configuration	MUS01, MUS05, MUS07, MUS13, HUS05, HUS06	Task creator
FR017	Decide which data owner can join the training on the fly.	Configuration	MUS01, MUS05, MUS13, HUS06	Task creator
FR018	Decide how to publish the trained model: publish to all users, a group of users or keep privately.	Configuration	MUS03, MUS04, MUS06, MUS08, HUS05, HUS10	Task creator
FR019	Start and end the training (either explicitly or by some implicit condition, e.g. a deadline until when participants can join, time budget, convergence criterion).	Usage	MUS01, MUS05, MUS07, MUS13, HUS04, HUS05, HUS06	Task creator
FR020	Cancel a task.	Usage	MUS01, MUS05, MUS07, MUS13, HUS04, HUS05, HUS06, HUS10	Task creator, Platform admin
FR021	Destroy a model.	Usage	MUS01, MUS05, MUS07, MUS13, HUS05, HUS06, HUS10	Task creator, Platform admin

FR022	Track the status of the tasks created by the user.	Usage	MUS01, MUS05, MUS07, MUS13, HUS04, HUS05, HUS06, HUS10	Task creator
FR023	Agree/disagree for new members to join or leave my task.	Configura-tion	MUS01, MUS05, MUS07, MUS13, HUS04, HUS05, HUS06, HUS10	Task creator
FR024	Participate in the training of that task's model.	Usage	MUS01, MUS04, MUS05, MUS07, HUS04, HUS05, HUS06, HUS10	Task mem-ber
FR025	Select datasets contributing to the task.	Usage	MUS01, MUS04, MUS05, MUS06, MUS07, MUS13, HUS04, HUS05, HUS06, HUS10	Task mem-ber
FR026	Have access to the trained model (in-termediate and/or final).	Usage	MUS01, MUS03, MUS04, MUS05, MUS07, MUS13, HUS04, HUS05, HUS06, HUS07, HUS10	Task mem-ber
FR027	Track the status of user tasks.	Usage	MUS01, MUS05, MUS07, MUS13, HUS04, HUS05, HUS06, HUS10	Task mem-ber
FR028	Be compensated for the data that the user contributed.	Usage	MUS02, MUS03, MUS4, MUS05, MUS07, MUS09, MUS10, HUS04, HUS05	Task mem-ber
FR029	Send a leave request to unjoin a ML task.	Usage	MUS01, MUS05, MUS07, MUS13,	Task mem-ber

			HUS05, HUS06, HUS10	
FR030	Use one or more trained models as they are published by task creators to the group.	Usage	MUS04, MUS05, MUS06, MUS07, HUS05, HUS06, HUS07, HUS10	Group owner
FR031	Add and remove users into the group.	Configuration	MUS01, MUS05, MUS07, MUS13, HUS09	Group owner
FR032	Create and launch synthetic tasks with “fake” users and synthetic data.	Usage	MUS01, MUS05, MUS07, HUS04, HUS05, HUS06	Researcher
FR033	Measure and compare performance of different Federated ML algorithms	Usage	MUS01, MUS05, MUS07, HUS03, HUS04, HUS05, HUS06, HUS07, HUS08, HUS10	Researcher
FR034	Create and delete users.	Configuration	MUS14, HUS09	Platform admin
FR035	Change the role / permissions / group of a user.	Configuration	MUS14, HUS09	Platform admin
FR036	Remove every type of information about all datasets (or the datasets themselves).	Usage	MUS14, HUS10	Platform admin
FR039	Track the status of all the tasks of the platform.	Usage	MUS14, HUS10	Platform admin
FR040	Grant another user admin privileges.	Configuration	MUS14, HUS09	Platform admin
FR041	Create / delete any jobs and datasets.	Usage	MUS14, HUS10	Platform admin
FR042	Configuration of privacy preserving data sharing methods	Configuration	MUS15, HUS10	Platform admin
FR043	Pre-processing data	Usage	MUS16, HUS04	Technical user

4.3.2 Non-functional requirements

Table 9 Non-functional requirements

ID	Description of the requirement	Category	Related User Stories
NR001	MUSKETEER platform should have a high availability	Availability	MUS01, MUS02, MUS03, MUS05, MUS06, MUS07, HUS03, HUS04, HUS05, HUS06
NR002	MUSKETEER should offer secure access and to be compliant with industrial security policies constraints	Security	MUS01, MUS14, HUS03, HUS04, HUS05, HUS06, HUS10
NR003	Automatic save custom modification of dataset	Recoverability	MUS01, MUS02, MUS03, HUS04, HUS05, HUS06
NR004	Automatic save custom modification of tasks	Recoverability	MUS04, MUS05, MUS07, HUS04, HUS05, HUS06
NR005	Continue task execution after a breakdown of the platform	Recoverability	ALL
NR006	MUSKETEER should be able to execute machine learning algorithms in a timely and efficient manner	Performance Efficiency	MUS01, MUS02, MUS03, MUS04, MUS05, MUS07, HUS01, HUS02, HUS04, HUS05, HUS06
NR007	MUSKETEER should enable the interconnection and exchange of information between data providers and the platform	Compatibility	MUS02, MUS03, MUS06, HUS03, HUS04, HUS05,

			HUS06, HUS07, HUS10
NR008	MUSKETEER should provide an easy-to-use and user-friendly interface in which the ML algorithms and visualization processes are supported by guides and manuals	Usability	ALL
NR009	Straightforward installation from end-user side	Usability	ALL
NR010	MUSKETEER should provide the appropriate logging mechanisms for all operations	Security	MUS01, MUS14, HUS10
NR011	MUSKETEER should provide the proper mechanisms for system upgrade with minimum downtime	Maintainability	ALL
NR013	MUSKETEER should be composed by independent components that are replaceable with minimum impact and effort	Portability	ALL
NR014	MUSKETEER should be to handle simultaneous requests on a timely and efficient manner	Maintainability	ALL
NR015	MUSKTEER should provide the mechanisms to recover after system failure conditions	Maintainability	ALL
NR016	MUSKETEER should handle software errors without affecting the platform overall functionality	Maintainability	ALL

4.4 MUSKETEER Technical Requirements

The previous section provided detailed descriptions of the functional and non-functional requirements that emerged during the requirement elicitation process. As a next step, technical requirements have been extracted from each of the aforementioned requirements individually and have subsequently been grouped into 37 requirements, as shown in the following table.

Table 10 Technical requirements

ID	Description of the requirement	Related functional and non-functional requirements	Scope

TR001	The MUSKETEER platform shall ensure that access control over datasets is applied according to the data policies and the terms of relevant active valid data sharing contracts.	FR001, FR002, NR002, NR0010	Security
TR002	The MUSKETEEER platform shall forbid unauthorised user access to the platform and the datasets.	FR001, FR002 NR002, NR0010	Security
TR003	The MUSKETEEER platform ensures different authorisation levels for accessing datasets.	FR001, FR002 NR002, NR0010	Security
TR005	MUSKETEEER end user must have a unique identification that will be used in all the data exchange/communications	FR001, FR002 NR002, NR0010	Security
TR006	Registration into the MUSKETEER Platform with username a password	FR001, FR002 NR010	Security
TR007	MUSKETEEER end user could create one or more new tasks	FR016, FR019	ML problem statement
TR008	In MUSKETEER a task must be defined as a problem statement that feeds from data and produces a trained machine learning model	FR003, FR007, FR008, FR010, FR011, FR013, FR014, FR015, FR016, FR019, FR025, FR026, FR030, NR006	ML problem statement
TR009	New tasks should obtain unique task identifier	NR008	ML problem statement
TR010	In MUSKETEER a task should have a general description	NR008	ML problem statement
TR011	The MUSKETEER Platform must, for each task, uniquely identify every input data from every end user	NR008	ML problem statement
TR012	Description of the input features. The meaning of every field must be explicitly described	NR008	ML problem statement
TR013	The MUSKETEER Platform must share a set of pre-processing algorithms such that every end user pre-processes its own raw data to obtain a common representation (e.g. high pass filtering, edge detection, bag of words with TFIDF weighting …)	FR043	ML problem statement

TR014	MUSKETEER pre-processing modules should always produce an output vector with the expected content and format	FR043, NR010	ML problem statement
TR015	MUSKETEER must make any ad hoc pre-processing algorithms (defined and implemented by the end users) shared with other users contributing to the task	FR043	ML problem statement
TR016	In MUSKETEER, for each task, definition and nature of the problem to be solved, must be shared among all the participants in a task such that they can contribute with new data to the training process.	NR008	ML problem statement
TR017	MUSKETEER Privacy Operation Modes must cover all kinds of privacy restrictions that end users would apply to his/her data	FR001, FR002 NR002, NR010	Privacy
TR018	Privacy restriction should be described in natural language to facilitate the specification of the task to the end user	FR001, FR002 NR002, NR010	Privacy
TR019	MUSKETEER Platform should envisage monetary rewards as well as collaborative results	FR028	Rewarding
TR020	Browsing for published active tasks by MUSKETEER end users	FR008, FR010, FR011	ML problem statement
TR021	Creation and/or access to published active tasks by MUSKETEER end users	FR013, FR014, FR015, FR016	ML problem statement
TR022	Running of the training procedure associated to a given MUSKETEER ML task	FR019, FR024, NR 006	ML problem statement
TR023	Monitoring of the progress of MUSKETEER ML tasks until completion	FR027	ML problem statement
TR024	MUSKETEER must provide the outcome of a task (reward, trained model, etc.)	FR018, FR028	ML problem statement
TR025	MUSKETEER must allow data to be transferred and joined either in the server or in a given user	FR006	Privacy
TR026	MUSKETEER must support the case where no raw data is transferred outside the client facilities (the ML model training must take place in the server by using the aggregated information from the clients)	FR042	Privacy

TR027	MUSKETEER must support the case where users want just to collaborate to create a predictive model without making data available	FR042	Privacy
TR028	Encryption methods should be supported in MUSKETEER for all those cases where data cannot be moved and the predictive model is private	NR010	Privacy
TR029	MUSKETEER server side should transform encrypted models among different private keys for all those cases where data owners use different private keys for homomorphic encryption	FR042	Privacy
TR030	MUSKETEER should offer private cloud storage for users' encrypted data	FR042	Privacy
TR031	MUSKETEER models should be defined so to have all information needed to operate (settings, cost function, prediction mode, gradient...)	FR016, FR019, FR025	Model Training
TR032	MUSKETEER models should be sent to any of the end users, such that it is possible to locally compute gradients or any other measurements	FR013, FR014, FR015, FR018, FR023, FR024, FR026	Model Training
TR033	MUSKETEER models should be sent to the server to combine them with the contributions from other users	FR013, FR014, FR015, FR018, FR026, FR030, NR007	Model Training
TR034	MUSKETEER model should not store any training data from the users without permission.	NR0010, FR042	Privacy
TR035	The ML training must take place in the server so to orchestrate all the steps to complete a given ML task.	FR016, FR017, FR019, FR030, NR007	Model Training
TR036	MUSKETEER server must to be able to send the model, wait for average gradients from end-users (client side) and update the model with the gradient information	FR013, FR014, FR015, FR018, FR026, NR007	Model Training
TR037	MUSKETEER server must to be able to receive in any moment the gradient of a model and update it.	FR018, FR026, NR007	Model Training

5 Conclusion

This deliverable D2.1 – Industrial and Technical Requirements presents the functional and non-functional end-user requirements and technical requirements for MUSKETEER platform derived from the needs of the two industrial scenarios considered within the project.

The scenarios were described and the user stories were formulated identifying the needs of different users in each scenario. The datasets provided by these scenarios have been characterised and analysed in order to identify the technical challenges implied with the processing of such datasets. The data flows in each environment were described and analysed in order to identify the most suitable privacy operation models that should be used in the MUSKETEER data platform.

Starting from domain-specific and business requirements listed, technical requirements elicitation was addressed so that they shall be met by the MUSKETEER Federated Machine Learning Architecture and so to ensure that the MUSKETEER IDP could be easily adapted and implemented in other relevant IDP and with a cross-sector dimension.

It is worth noticing that a revision of the technical requirements will be done in M22.

6 References

DICOM Standard: https://www.dicomstandard.org/

HL7 Standard: http://www.hl7.org/

ACR - American College of Radiology, Prostate Imaging Reporting & Data System - PI-RADS 2015 Version 2, full text document: https://www.acr.org/-/media/ACR/Files/RADS/Pi-RADS/PIRADS-V2.pdf?la=en

TCIA collections: https://www.cancerimagingarchive.net/nbia-search/?CollectionCriteria=PROSTATE-MRI

Zenodo collections: https://zenodo.org/search?page=1&size=20&q=keywords:%22prostate%20cancer%22,%20%22MRI%22&type=dataset

OpenAIRE collections: https://explore.openaire.eu/search/find/datasets?instancetypename=%22Dataset%22&keyword=prostate%20MRI

Radiology Assistant site: http://www.radiologyassistant.nl/en/p59987056acbb4/prostate-cancer-pi-rads-v2.html

Serope Kalpajian, Steven R. Schmid (2013) -Manufacturing Engineering and Technology- Pearson 7th Edition

Menachem Kimchi, David H. Phillips (2017) -Resistance Spot Welding: Fundamental and Application for the Automotive Industry- Morgan & Claypool

Industrial Data Space, IDS Reference Architecture Model (2018)

Framework for the IDS Certification Scheme, Version 2, IDS Reference Architecture Model (2018)